PB86-107638

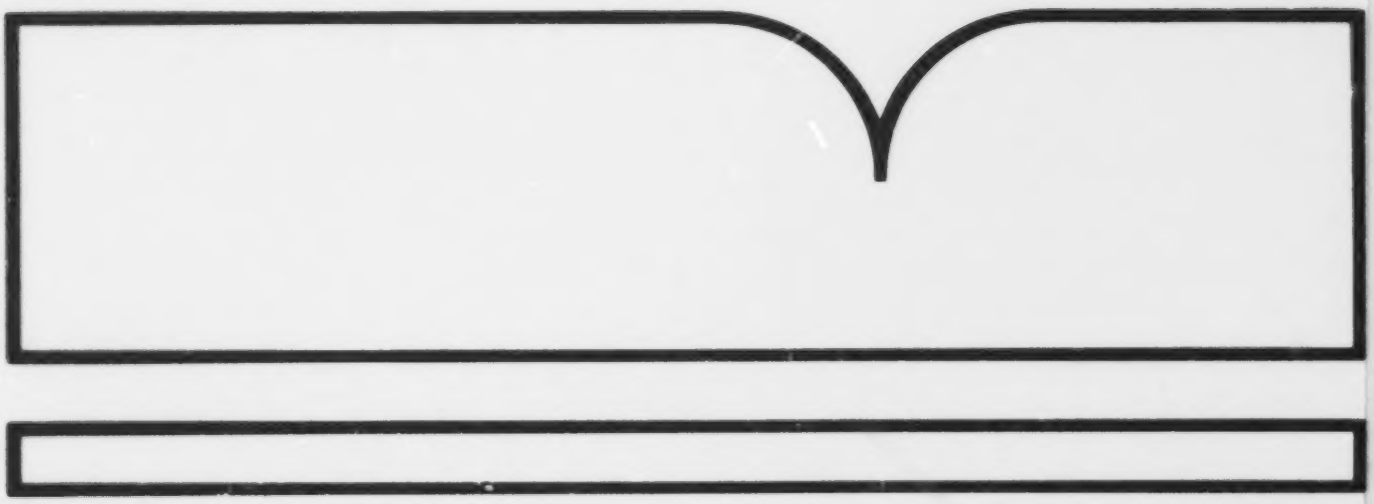
# Receptor Model Technical Series. Volume 6 A Guide to the Use of Factor Analysis and Multiple Regression (FA/MR) Techniques in Source Apportionment

New York Univ. Medical Center, NY

Prepared for

Environmental Protection Agency
Research Triangle Park, NC

Jul 85



U.S. Department of Commerce
National Technical Information Service
NTIS

PB86-107638

United States
Environmental Protection
Agency

Office of Air Quality
Planning and Standards
Research Triangle Park NC 27711

EPA-450/4-85-007
July 1985

Air

EPA

# Receptor Model Technical Series VI: A Guide To The Use Of Factor Analysis And Multiple Regression (FA/MR) Techniques In Source Apportionment

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

**TECHNICAL REPORT DATA**
*(Please read Instructions on the reverse before completing)*

| 1. REPORT NO. | 2. | 3. RECIPIENT'S ACCESSION NO. |
|---|---|---|
| EPA-450/4-85-007 | | PB86 107638 |

| 4. TITLE AND SUBTITLE | 5. REPORT DATE |
|---|---|
| Receptor Model Technical Series, Vol. VI: Factor Analysis And Multiple Regression (FA/MR) Techniques In Source Apportionment | July 1985 |
| | 6. PERFORMING ORGANIZATION CODE |
| 7. AUTHOR(S) | 8. PERFORMING ORGANIZATION REPORT NO. |
| Paul J. Lioy, Theo J. Kneip and Joan M. Daisey | |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS | 10. PROGRAM ELEMENT NO. |
| Institute Of Environmental Medicine<br>NYU Medical Center<br>550 First Avenue<br>New York, NY 10016 | 11. CONTRACT/GRANT NO.<br>4D2975NASA |
| 12. SPONSORING AGENCY NAME AND ADDRESS | 13. TYPE OF REPORT AND PERIOD COVERED |
| U. S. Environmental Protection Agency<br>OAQPS, MDAD, MD 14<br>Research Triangle, NC 27711 | 14. SPONSORING AGENCY CODE |

15. SUPPLEMENTARY NOTES

EPA Project Officer: Thompson G. Pace, III

16. ABSTRACT

The anticipated change in the form of the particulate matter standard from total suspended particulate to matter with aerodynamic diameter equal to or less than 10 micrometers ($PM_{10}$) will require, in some instances, more sophisticated approaches to identifying primary sources of $PM_{10}$. This is the sixth document in a series of user oriented receptor modeling guidance.

Over the past twelve years, a number of multivariate methods have been used to determine the sources of mass emitted in a number of cities. This document focuses primarily on the FA/MR technique. However, the procedures required to identify potential tracers or source profiles and to validate the results are applicable to all

17. KEY WORDS AND DOCUMENT ANALYSIS

| a. DESCRIPTORS | b. IDENTIFIERS/OPEN ENDED TERMS | c. COSATI Field/Group |
|---|---|---|
| Receptor Modeling<br>Factor Analysis<br>Multiple Regression<br>Multiple Linear Regression<br>Statistical Inference<br>Source Apportionment | | |

| 18. DISTRIBUTION STATEMENT | 19. SECURITY CLASS (This Report) | 21. NO. OF PAGES |
|---|---|---|
| | | 110 |
| | 20. SECURITY CLASS (This page) | 22. PRICE |

EPA Form 2220-1 (Rev. 4-77) PREVIOUS EDITION IS OBSOLETE

EPA-450/4-85-007

# Receptor Model Technique Series IV: A Guide To The Use Of Factor Analysis And Multiple Regression (FA/MR) Techniques In Source Apportionment

By

Paul J. Lioy, Theo. J. Kneip, And Joan M. Daisey
Institute of Environmental Medicine
New York University Medical Center

Contract No. 4D2975NASA

EPA Project Officer:
Thompson G. Pace

U.S. ENVIRONMENTAL PROTECTION AGENCY
Office Of Air And Radiation
Office Of Air Quality Planning And Standards
Research Triangle Park, NC 27711

July 1985

This report has been reviewed by the Office Of Air Quality Planning And Standards, U.S. Environmental Protection Agency, and approved for publication as received from the contractor. Approval does not signify that the contents necessarily reflect the views and policies of the Agency, neither does mention of trade names or commercial products constitute endorsement or recommendation for use.

EPA-450/4-85-007

## PREFACE

Receptor Model Technical Series

Volume VI

A Guide to the Use of Factor Analysis and Multiple Regression Techniques in Source Apportionment

In order to meet the requirements of the 1977 Clean Air Act regarding attainment of the National Ambient Air Quality Standards for particulate matter, EPA has been preparing guidelines for use in identifying and quantifying source contributions to measure ambient particulate matter concentrations. Many analysis techniques and models have been developed for the purpose of source apportionment. Receptor models are those that are based primarily on ambient concentration data gathered at the receptor, and are used to determine the sources contributing mass at the site.

Guidance for using source apportionment techniques has been compiled by EPA into the Receptor Model Technical Series. The first four volumes in the Technical Series have primarily addressed receptor model source apportionment techniques. Volume I (EPP-450/4-81-016a), entitled "Overview of Receptor Model Application to Particulate Source Apportionment", introduces the concept of receptor models and briefly discuses the various types of receptor models and briefly discusses the various types of receptor models and their applications. Volume II (EPA-450/4-81-016b) pertains to the "Chemical Mass Balance" model and provides

information on model theory, data requirements and case studies of the application of the model to emission control strategy development. Volume III (EPA-450/4-83-014), the "User's Manual for (a) Chemical Mass Balance Model", documents a computer program that performs source apportionment using the weighted least squares and other optional forms of the mass balance equations. The user's guide provides a complete program listing, an example set of input and output data, and further discussion of model theory and use.

Volume IV of the series (EPA-450/4-83-018), "Summary of Particle Identification Techniques" gives an overview of the methods and equipment generally used in particle characterization for source apportionment studies. The discussion includes sampling and analytical methods, choice of filter media, particle properties and source fingerprints, costing and method selection criteria.

Volume V (EPA-450/4-84-020), "Source Apportionment Techniques and Considerations in Combining Their Use", provides guidance for the coordinated use of the various receptor and source model techniques in source apportionment activities. Summary discussions of the available receptor and source models are presented. The use of the models is discussed in a phased approach starting with analyses of low complexity and cost and proceeding to analyses of greater complexity and cost which produce more quantitative results. Input data requirements for each phase and example case histories are provided.

The present volume (VI), "A Guide to the Use of Factor Analysis and Multiple Regression (FA/MR) Techniques is Source Apportionment", provides an informative analysis of the theory and application of the combined technique of factor analysis and multiple regression to identify sources and estimate their contributions. Features of the volume are a thorough discussion of the methods for applying these statistical techniques to a large data base, interpretation of the results and techniques for validation of the tracers and regression coefficients. The types of information and tests available to determine the stability of factor analytical solutions, and the completeness of a regression analysis, are also described in the text.

## Abstract

### Receptor Model Technical Series

### Volume VI

### A Guide to the Use of Factor Analysis and Multiple Regression (FA/MR) Techniques in Source Apportionment

One of the major requirements of the Clean Air Act is to attain the National Ambient Air Quality Standard for particulate matter. In addition, with the anticipated changed in the form of the standard from total suspended particulate matter to a standard for matter with an aerodynamic diameter of $\leq$ 10 μm (PM-10), more sophisticated approaches to identifying the primary sources of PM-10 will be required in some instances. The present document is the sixth in the user oriented receptor modeling series. Over the past twelve years, a number of multivariate methods have been used to determine the sources of mass emitted in a number of cities. The present document focusses primarily on the FA/MR technique; however, the procedures required to identify potential tracers or source profiles, and validate the results are applicable to all.

The specific items covered in the text are 1) the definition and use of factor analytical modeling techniques; 2) discussion and analysis of the results of applications of the factor analytical technique to

real data sets, including attempts for validation, 3) the definition of the regression model, and 4) the application of the regression technique to apportion the particulate mass based upon tracer selection criteria established for FA. Finally, the important task of validating the source contributions obtained from a stepwise regression is examined by examples, and weaknesses discussed. To assist the reader, the results from a complete FA/MR analysis are followed in the application section of both the Factor Analysis and Multiple Regression Chapters.

Other multivariate techniques used in source apportionment are identified in Section 4.0 of the report. No critical review or detailed discussion is provided for these. Further examination of the previous applications to ambient data sets are left to the reader.

Acknowledgements

The authors wish to thank the Project Officer, Thompson G. Pace, for providing comments and criticism throughout the development of this document. Ms. Laraine Wittrup and Mrs. Betty McCarthy for typing the manuscript, and Mrs. Mary Jean Yonone Lioy for editing the final version. Further appreciation is extended to the technical reviewers of the document, Dr. Thomas Dzuby, Dr. Charles Pratt and Mr. William Cox of the U.S. EPA, and Mr. Stuart Dattner of the Texas Air Control Board.

Table of Contents


| | Page |
|---|---|
| 1.0 Introduction | 1. |
| 1.1 Background | 2. |
| 2.0 Factor Analysis | 7. |
| 2.1 The Factor Analytical Model | 11. |
| 2.2 Selection of Variables | 22. |
| 2.3 Commonly Available Statistical Software | 27. |
| 2.4 Application of Factor Analysis to Air Pollution - Data and Interpretation of Results | 32. |
| 2.4.1 General Considerations | 32. |
| 2.4.2 Preliminary Examination of Data | 33. |
| 2.4.3 Factor Analysis Solutions | 34. |
| 2.4.4 Type of Factors | 38. |
| 2.5 An Example of Application to Air Pollution Data | 40. |
| 2.6 Validation of Factor Analysis | 45. |
| 2.6.1 Introduction | 45. |
| 2.6.2 Source Composition Profiles | 46. |
| 2.6.3 Factor Validation | 47. |
| 2.6.4 Evaluation of Source Inventories | 48. |
| 2.6.5 Factor Stability | 49. |
| 2.7 References | 52. |
| 3.0 Multiple Regression Analysis | 57. |
| 3.1 Simple Bivariate Linear Regression Model | 58. |

3.2 Multiple Regression Analysis 63.

3.3 An Example of An Application of Multiple Regression Analysis to Air Pollution Data 66.

3.4 Regression Model Validation 71.

3.4.1 Regression Coefficient Validation 73.

3.4.2 Meteorological Relationships 80.

3.4.3 General Model Evaluation 81.

3.4.4 Summary 84.

3.5 Interpretation 85.

3.5.1 Summary 88.

3.6 References 89.

4.0 Appendix: Alternative Approaches to Regression Analysis 93.

A.1 Target Transformation Factor Analysis (TTFA) 93.

A.2 Multiple Linear Regression on Tracers/Factor Analysis [MCR(T)/FA] 95.

A.3 Regression of Absolute Principal Component Scores 96.

A.4 References 99.

## 1.0 INTRODUCTION

A primary purpose of the receptor modeling series, Volumes I through V, is the transfer of information and techniques that are extremely useful to the research community, to environmental specialists and to scientists in air pollution regulatory agencies (1-5). Subsequently, these individuals can use the techniques to examine air pollution in locations with high particulate matter concentrations, toxic pollutant emissions from multiple sources, visibility degradation, acid deposition, and hazardous organic pollutants. No single receptor modeling technique is always applicable or desirable for all situations. Therefore, a judgement must be made on the utility of a particular technique or group of techniques for the problem to be studied. In this regard, a preliminary evaluation must be conducted to determine the quality and quantity of available, retrievable input data, and ancillary information proposed for use in a model.

In Volume VI, the application of multivariate techniques to receptor modeling will be examined. There will be an emphasis on common factor analysis and stepwise regression analysis and on application to and interpretation of problems associated with particulate matter pollution (irrespective of any particle cut size). The Factor Analysis/Multiple Regression (FA/MR) Receptor Model, the topic of this volume, first uses factor analysis to identify sources of particulate matter and to select source emission tracers. Stepwise multiple regression analysis is then used to obtain a quantitative relationship between the source tracers and particle mass concentration. However, both are independent mathematical techniques. The purpose of volume VI is to familiarize air

pollution managers with the framework and approach necessary to use this receptor modeling technique. It is not intended as a treatise on the statistical and mathematical aspects of each topic. The volume will provide such individuals with enough information to decide if the technique will be useful in solving a given problem.

## 1.1 Background

Because of the nature of atmospheric processes and meteorology, the concentrations of individual air pollutants will often vary simultaneously. This occurs irrespective of the sources; therefore, it is difficult to resolve tracers or markers emitted from individual sources. Frequently, the difficulties in differentiating individual sources are associated with the simultaneous increase and decrease (intercorrelation) of the selected variables, which are usually observable on time series plots (Figure 1.1 and 1.2). In most applications of FA/MR to particulate matter, the variables are the measured concentrations of trace elements; for example, vanadium (V), nickel (Ni), lead (Pb), and cadmium (Cd).

In Figure 1.1, the two elements portrayed (A and B) increase and decrease simultaneously in almost all cases, indicating an almost perfect correlation coefficient of $r = 1$ (an actual value of $r = 0.99$ in this case). In Figure 1.2, real data for nickel and vanadium are shown which have a very high correlation of $r = 0.84$ with the most dramatic variations occurring during pollution episode periods. These are identified by the shaded area. Other variables shown, Pb and Cd, are not as firmly coupled to the previous two, having $r = 0.62$ and 0.39, and 0.74 and 0.54, with V and Ni respectively. Both Pb and Cd still show major

Figure 1.2

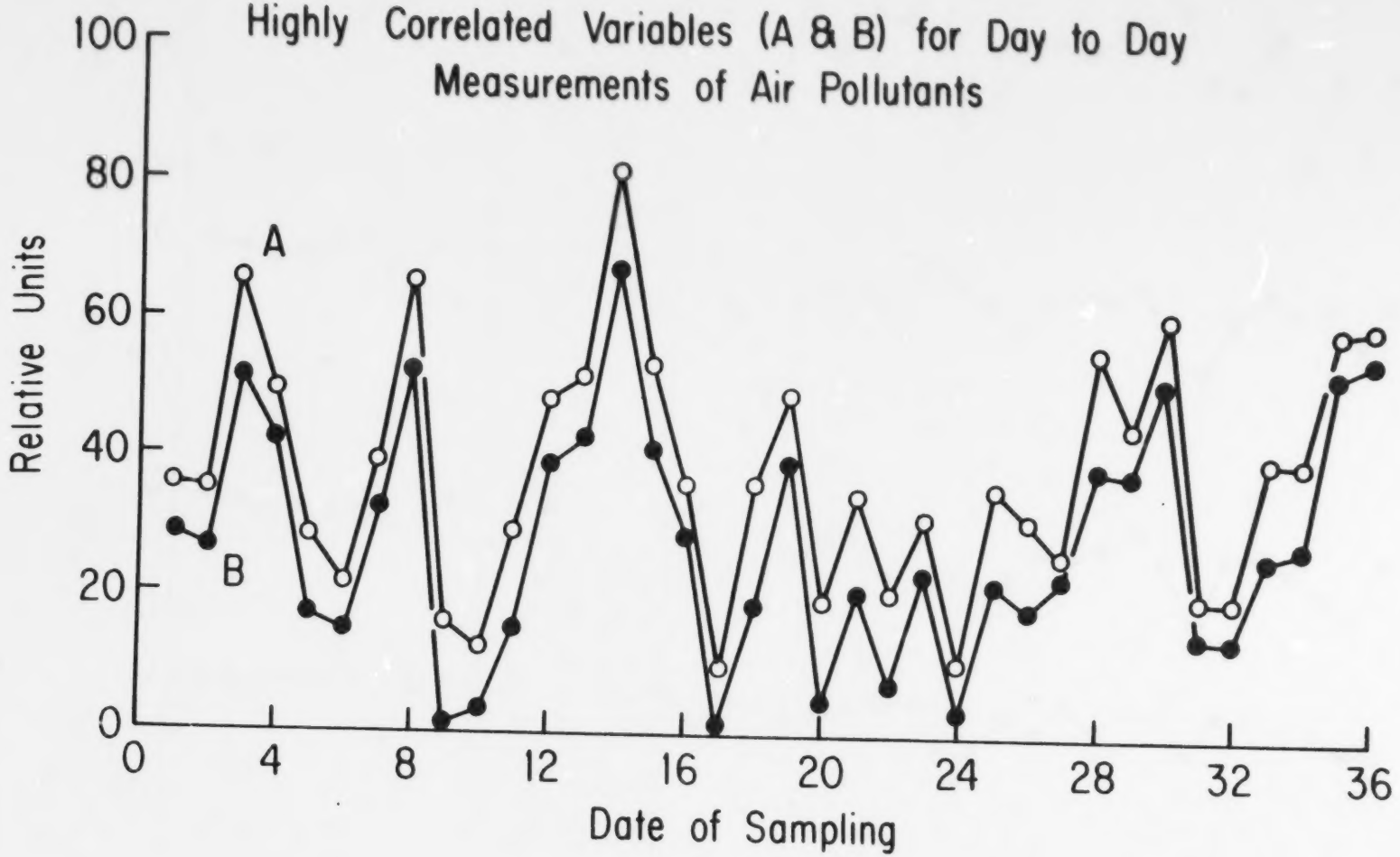
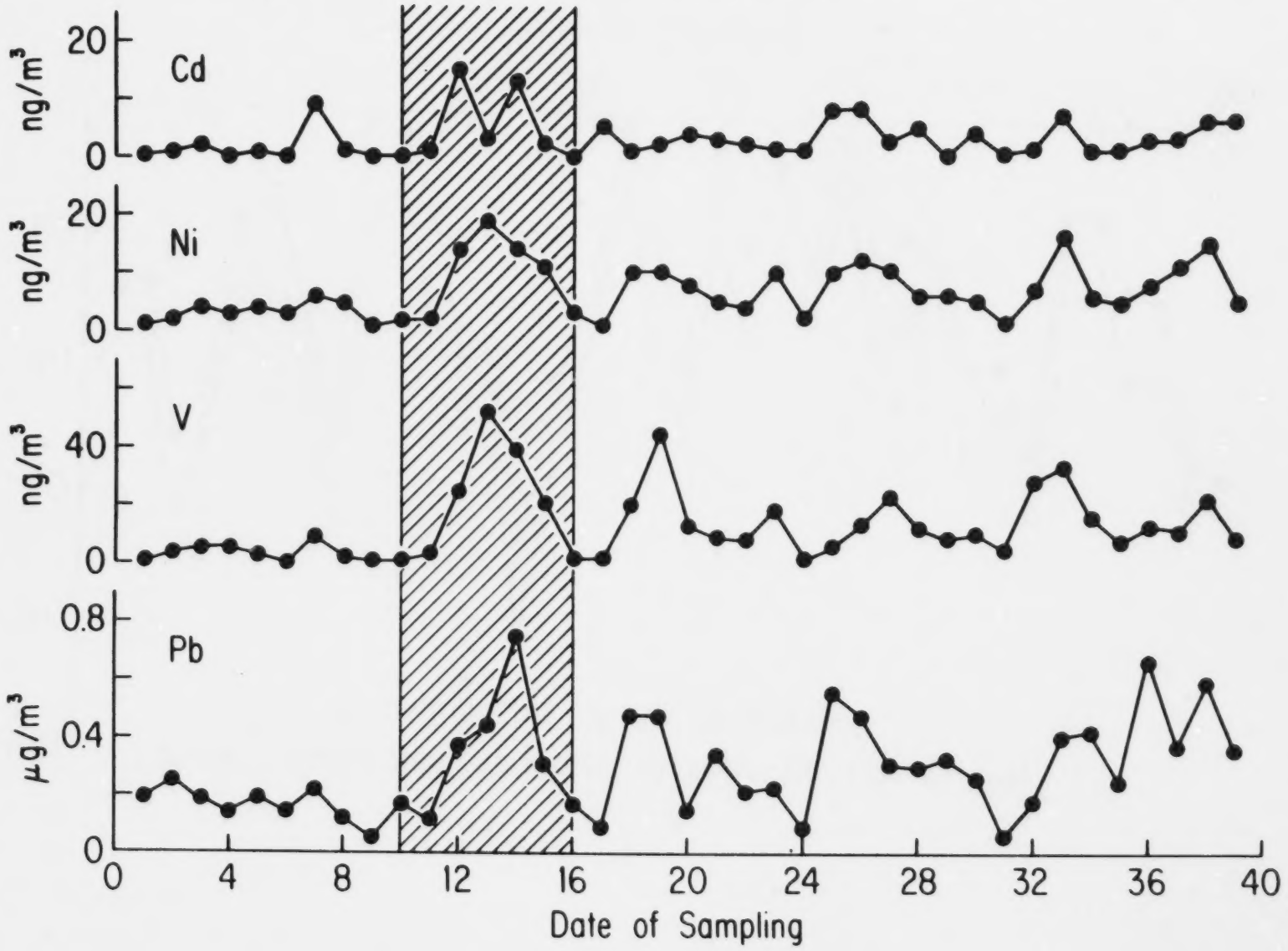

Highly Correlated Variables (A & B) for Day to Day Measurements of Air Pollutants
Relative Units
100
80
60
40
20
0
A
B
0
4
8
12
16
20
24
28
32
36
Date of Sampling

Figure 1.1 Data Collected for Lead, Cadmium, Vanadium, and Nickel on Thirty nine Sampling Days in Elizabeth, N.J.



concentration excursions during an episode.

This example, however, does not immediately suggest that individual sources or source categories contributing to the particulate mass can be identified from the data and suggests the need for statistical techniques to unravel sources. In the past, attempts have been made to use co-varying materials as tracers for the sources of the particulate mass in linear regression models. Errors in source assignments are inevitable in such a case. Without incorporation of independent selection procedures to verify the acceptability of an element as a tracer the models were usually inadequate. For V and Pb in the example given in Figure 1.2, this is not readily apparent, even though these elements are usually tracers for different sources, i.e., fuel oil and automobiles, respectively.

One of the major problems with air pollution data is that the underlying reason for the simultaneous variation of some of the tracers is the meteorology. This can affect the accumulation of individual or all particulate species and vapor phase species. Frequently, the effect is identified by sudden increases in particulate mass and species from one day to another, a peak day or a series of high concentration days, and then a decline to much lower concentrations for a day or a protracted period of time (6). At this point, it could be suggested that it is inappropriate to attempt a mass apportionment for a receptor site by multivariate techniques. However, within the data, patterns are present, and in some cases these can be inferred by the lack of uniform rates of change in concentration from day to day for individual species (such as for V and Pb in Figure 1.2). Unfortunately, such qualitative

features alone are not usually sufficient to begin to devise source-receptor relationships and to allocate mass contributions from sources. The latter requires the use of specific multivariate techniques to disentangle the co-variation of a subgroup of potentially useful elemental source tracers.

Use of multivariate techniques requires a priori 1) measurement of a large number of tracers (elements, etc.), and 2) large numbers of data observations or samples. The latter are usually referred to as cases, and represent a specific time period of sampling (1 hr, 4 hr, 24 hr, etc.). This point is extremely important since the financial and personnel resources must be allocated to collect a sufficient number of samples for analysis or a sufficient number of archived samples must be available for chemical composition analyses. In any case, the samples to be used must be obtained from a single site in the area under consideration in the receptor modeling study. Two sites located near each other can be used most effectively in source apportionment validation studies and in detection of differences due to localized emission sources (e.g., minor emission sources, < 50 T/yr). The latter would be especially useful for toxic substance investigations since many of these sources will not necessarily appear on an emissions inventory.

A further caveat is that multivariate techniques are only mathematical tools which help interpret the data, and are only as good as the data entered (qualitatively and quantitatively). The latter subject will be addressed in another section; however, it can not be overstated since solutions to a poorly constructed data set would ultimately be of no value to an environmental specialist and eventually to the regulator.

## 2.0 FACTOR ANALYSIS

Factor analysis (7) comes from the field of social science and has been described by Rummel (8) as the "calculus of the social sciences". It constructs a model which mathematically describes any behavioral relationships that can be deduced or predicted from the specific phenomena under investigation. In the case of air pollution, the model describes source emissions relationships from the characteristic changes in pollutant species and chemical element concentrations. The mathematics required include correlation matrices, diagonalization of axes, eigenvalues, eigenvectors and principal axis rotations with the computations normally performed by computer. The details of the mathematical formulations will not be subject of this document. However, the basic concepts will be reviewed.

Over the past ten years, factor analysis has been a useful tool in the source apportionment of particulate matter (9-16). The factor models developed for source apportionment have been used to infer emission source tracer relationships and subsequently in the selection of individual elements or chemical species for use in the regression source apportionment analyses (9-13,17). Recently, Thurston (16) has used the results of principal component analysis to obtain estimates of the actual source emissions composition profiles. However, this method requires more a priori information about the completeness of the source tracer list than is required by the more general FA/MR technique.

The technique of factor analysis groups the selected variables according to their common variations. These groupings are called factors. In air pollution, the variables are normally the source emission

tracers and these tracers are normally elements or ions. Experience has shown that in order to conduct a successful factor analysis to identify the major sources of particulate pollutants, the investigator should attempt to obtain at least two distinct potential tracers per source type. Since two may not be available in all cases, one marker could be used nominally to represent a possible source. The investigator may wish to use as many potential markers as are available. It should be emphasized, however, that the use of many elements to characterize a particular source type will not necessarily yield a better result. This is because factor analysis attempts to group the variation of the elements provided. In the case where there are many elements for one source type, this can lead to indications that that source type dominate the variation of the data (which is a bias). A balance should be sought in the number of tracers to be used as input to the factor analysis study and the number to be used to characterize an individual source type, as is explained in a subsequent section (17).

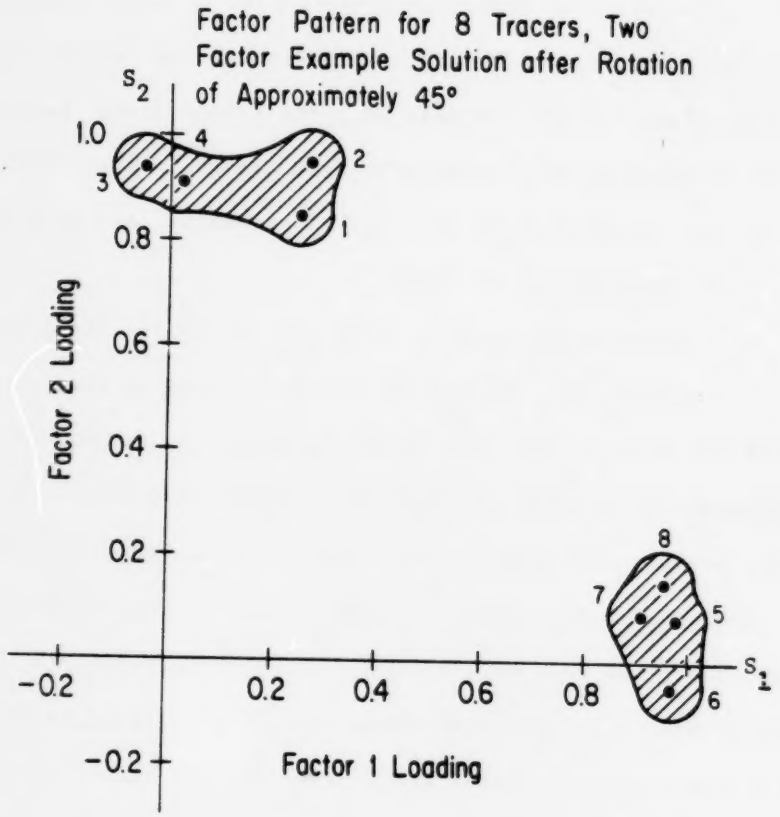
A major advantage of using common factor analysis is that it takes "n" tracers that are presumed to be tracers related to the air pollution receptor site under investigation and groups these "n" tracers in such a way that "m" new independent tracers are constructed as factors. These describe the variation of the data based upon the clustering of the original variations. To be an effective tool "m" needs to be less than "n", which indicates that the dimension of the air pollution problem has been reduced into linear combinations of the tracers called factors. This procedure is common to all scientific theory and is known as the principal of <u>parsimony</u> (8).

For factor analysis, the reduction in the number of tracers is associated with the size (rank) of the variable correlation matrix (degree of association among each of the variables under consideration). This reduction in size is related to mathematical constraints, which are found under the heading of communality considerations* (7). Factors constructed in any factor analysis model will have "n" dimensions, but can be represented by geometric plots in two or three dimensions. Computer programs allow for the examination of the factors as a series of two dimensional plots. In this way, the investigator can construct a visual interpretation of the model from the analytical solutions, obtain qualitative information on clusters of variables, and gain further ideas for factor analyses of the data.

As a preliminary example, in Figure 2.1 we see the grouping of elemental tracers (E1 through E8) along the axes of what for now will be called two factors (S1, S2). These groupings are identified by values of loadings (correlations with factors) greater than 0.90 on either factor. It can be seen that there is a definite separation of the elements with two groups forming along the individual new rotated factor axes. The values along the axes are the correlation coefficients of the original tracers with the factors; these are always defined as the loadings of the element with the factor represented by that axis. For this rather straightforward example, the split among the elemental tracers is quite explicit, and the elements E1 to E4 and E5 to E8 are related to factors

*The communality of a given variable is the sum of the squares of the loadings (correlations) of a given variable with each factor, and is determined by an iterative process in classical factor analysis. The estimated communality for any variable must not exceed 1.0 (7).

Figure 2.1


Factor Pattern for 8 Tracers, Two Factor Example Solution after Rotation of Approximately 45°
$S_2$
1.0
0.8
0.6
0.4
0.2
-0.2
-0.2
0.2
0.4
0.6
0.8
$S_1$
Factor 2 Loading
Factor 1 Loading
1
2
3
4
5
6
7
8

2 and 1 respectively. This result indicates that two new independent factors are adequate to represent the original eight elemental tracers. In a factor analytical solution of pollution data, the eight tracers should be individual source tracer elements and the two factors would represent individual source types. In general terms, the above example can be classified as a two factor model which has combined the variation of eight tracers into the two new distinct factors.

A number of steps must be completed to obtain a factor analytical solution and yield sufficient information from the data to eventually interpret the factors. An outline of the general framework of the mathematics used in factor analysis is presented in the next section, with the main focus being an understanding of the model design. The details are found in an excellent book by Harman (7) which includes numerous problems.

## 2.1 THE FACTOR ANALYTICAL MODEL

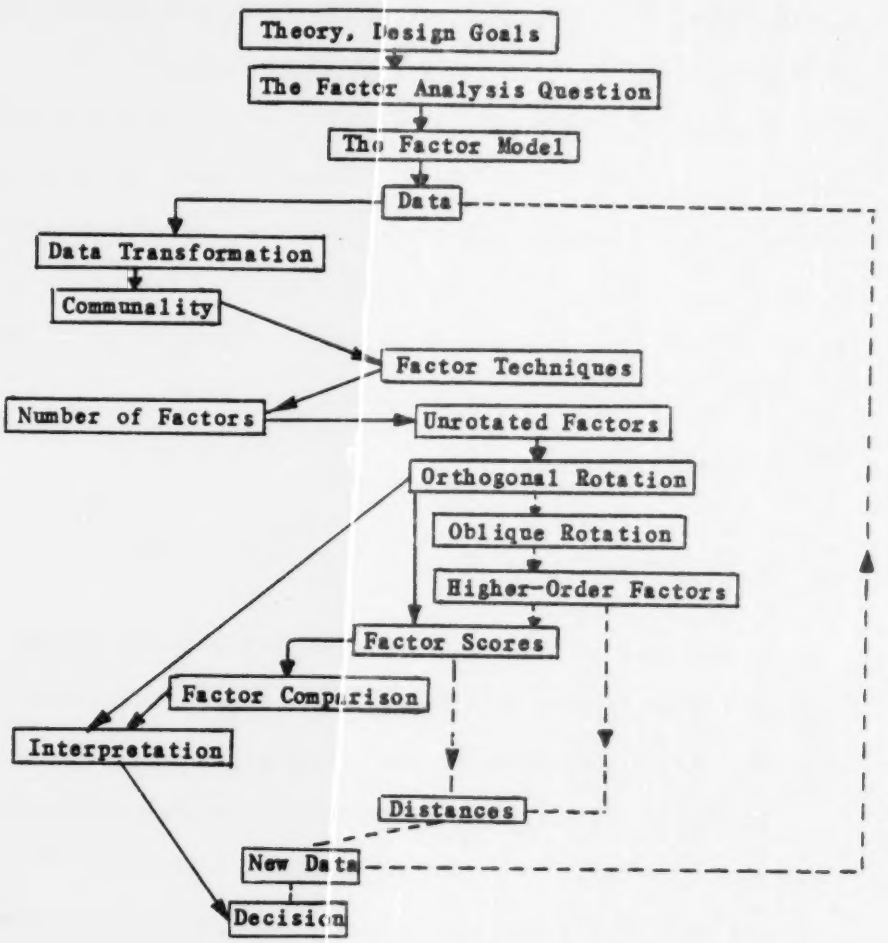
A very concise flow diagram of the factor model design is shown in Figure 2.2 as adapted from Rummel (8). Frequent reference to this figure will aid in understanding the remainder of this Chapter.

Basically there are two different factor analytical models: 1) Common Factor Analysis; 2) Principal Component Analysis.

Variants of the latter exist and are used in computer program applications. Although the emphasis will be on Common Factor Analysis, the Principal Component Analytical Model will be discussed first. It involves a number of simplifying assumptions with respect to the variables set, but the solutions are exact.

Figure 2.2

FACTOR ANALYSIS RESEARCH DESIGN FLOW DIAGRAM




KEY —> = usual flow of factor analysis design; ---> alternative flow

Adapted from Rummel (8)

The principal component model is defined by the linear equation:

$$z_i = a_{i1}F_1 + a_{i2}F_2 + \cdots + a_{in}F_n \qquad \text{Eq. 1.}$$

$(i = 1 \text{ to } n)$

where the original n variables (tracers) are distributed as linear combinations to produce the new uncorrelated variables called principal components $F_1$, $F_2$, $F_n$ (i.e., the number of factors equals the number of variables), and the $a_{ij}$'s are the loadings on the factors. $z_i$ is called the standardized variable of observed variable $X_i$, such as the elements Vanadium or Zinc, and for individual cases is of the form:

$$z_{ip} = \frac{X_{ip} - \bar{X}_i}{\sigma_i} \qquad \text{Eq. 2}$$

where: $X_{ip}$ = pth observation of the ith variable

$\bar{X}_i$ = variable mean

p = variable mean observations

The main advantage of using z , which is called a z-score, is that it has a mean of zero and a standard deviation of 1, which simply means that only the relative deviations of the original variables are retained. The obvious advantages of having all variables in this form are the elimination of problems associated with 1) different variable scales, and 2) the differing ranges of the measurements observed for each variable (7).

For principal component analysis, an important property is that the individual factors (or components) attain a maximum contribution from the variances of all the selected variables (tracers). In the present

analysis this involves reproducing the original correlation matrix for the (n) tracers used to develop the factor models, assuming the tracers selected describe the air pollution problem completely, assuming that there are no errors, and assuming the tracers selected are the most representative of an individual source profile. This feature can prove to be beneficial in cases where the emission sources are known (16) since the factor solution can be scaled to give new linear combinations called factor scores. These scaled or normalized equations can then be adjusted to estimate source contributions. The equation for the factor score is of the form:

$$f_j = S_{j1}z_1 \; S_{j2}z_2 + \cdots S_{jn}z_n \qquad \text{Eq. 3.}$$

where: $f_j$ = the factor score,

$S_{ji}$ is the factor score coefficient,

$z_i$ is the normalized or standardized variable, (as defined in Eq (1), i = 1 to n.

In contrast, common factor analysis attempts to maximize the common variation among the available variables and produce correlations which approximate the original correlation matrix of the original variables (a reproduced correlation matrix). More simply, it maximizes the shared variation of the available variables. The equation describing the common factor model is:

$$z_i = a_{i1}F_1 + a_{i2}F_2 + \cdots + a_{im}F_m + \mu_i Y_i \quad i = 1 \text{ to } n \qquad \text{Eq. 4.}$$

where each of the n observed variables used in the model are described by m common factors (m < n) and a unique factor ($Y_i$). The $a_{ij}$ and $\mu_i$ which are called loadings, are the correlations of the jth variable with the mth factor. The unique factor ($Y_i$) will include the remaining variance of the variable $z_i$. This latter portion of the equation describes

the part of a variable which is uncorrelated with the other variables or the derived m factors, and may suggest the need for more tracers if the $\mu_i$ is large. A generalized print-out from this model is as shown in Table 2.1.

The ccmmon factor analysis, therefore, provides a check of the completeness of the tracer set used to develop the model. In fact, one test for the applicability of common factor analysis was developed by Guttman (18). He indicated that the original correlation matrix (R) should be inverted ($R_{nxn}^{-1}$) and the off diagonal elements of the matrix should approach zero. If these are not near zero, more variables (tracers) may be required to describe the features of the phenomenon (air pollution) under investigation. It should be noted, however, that a number of other conditions must also be satisfied (7).

To summarize, a basic difference between common factor analysis and principal component analysis is that the common factor analysis attempts to separate common, specific (not associated with included variables) and random error variances of the selected variable list, whereas the latter assumes the availability of a complete set of variables, which is most often not the case for air pollution problems. The results of the two analyses will become nearly equivalent as $\mu^2_i$ approaches zero. However, because the initial conditions used to extract the factors from PCA and FA are different, the results will not achieve exact convergence.

An example of the matrix that can be obtained as a result of a factor analysis is shown in Table 2.2. This could be for any city in the U.S.A., but it happens to be the result of a common factor analysis for

TABLE 2.1

INFORMATION ARRAY FROM A COMMON FIVE FACTOR

ANALYSIS WITH SEVEN ELEMENTAL TRACERS

| Tracers | Factor 1 | Factor 2 | Factor 3 | Factor 4 | Factor 5 | Factor Y |
|---|---|---|---|---|---|---|
| E1 | $a_{11}$ | $a_{12}$ | $a_{13}$ | $a_{14}$ | $a_{15}$ | $\mu_{m1}$ |
| E2 | $a_{21}$ | $a_{22}$ | $a_{23}$ | $a_{24}$ | $a_{25}$ | $\mu_{m2}$ |
| E3 | $a_{31}$ | $a_{32}$ | $a_{33}$ | $a_{34}$ | $a_{35}$ | $\mu_{m3}$ |
| E4 | $a_{41}$ | $a_{42}$ | $a_{43}$ | $a_{44}$ | $a_{45}$ | $\mu_{m4}$ |
| E5 | $a_{51}$ | $a_{52}$ | $a_{53}$ | $a_{54}$ | $a_{55}$ | $\mu_{m5}$ |
| E6 | $a_{61}$ | $a_{62}$ | $a_{63}$ | $a_{64}$ | $a_{65}$ | $\mu_{m6}$ |
| E7 | $a_{71}$ | $a_{72}$ | $a_{73}$ | $a_{74}$ | $a_{75}$ | $\mu_{m7}$ |

*calculated by $\mu^2_{mi} = 1 - a^2_{11} - a^2_{12} - a^2_{13} - a^2_{14} - a^2_{15}$

Table 2.2.

ROTATED FACTOR LOADING (PATTERN) FOR SUMMERTIME PARTICULATE MATTER MEASUREMENTS IN CAMDEN, N.J.

| Variable | Factor 1 | Factor 2 | Factor 3 | Factor 4 | Factor 5 |
|---|---|---|---|---|---|
| Pb | -0.050 | .192 | .184 | .719 | .242 |
| Mn | .107 | .825 | .094 | .051 | .000 |
| Cu | .118 | .097 | .164 | .143 | .700 |
| V | .802 | .151 | .092 | -.039 | .050 |
| Cd | -.029 | -.013 | .674 | .557 | -.080 |
| Zn | .050 | .118 | .661 | .083 | .286 |
| Fe | .274 | .608 | .017 | .231 | .303 |
| Ni | .879 | .342 | .097 | .032 | .073 |
| $IPSO_4$ | .598 | -.023 | -.083 | .010 | .076 |

summertime particulate matter data in Camden, N.J. (19) The intent here is not to interpret the data, but to explain basic information available from the analysis. The table shows the factors from a common factor analysis of n = 9 tracers and the resultant m = 5 factors. Each factor, $F_1$ to $F_5$ is composed of a linear combination of loadings which are related to the individual tracers. When placed in the form of the original equation, the variation of lead is described by the rows of Table 2.2:

$$Pb = a_{11}F_1 + a_{12}F_2 + a_{13}F_3 \ldots a_{1m}F_m + (\mu_{Pb}Y_{pb}) \text{ or more specifically:}$$

$$Pb = -0.05\ F_1 + 0.192\ F_2 + 0.184\ F_3 + 0.719\ F_4 + 0.242\ F_5 + \mu_{Pb}Y_{Pb} \qquad \text{Eq. 5.}$$

If this set of factors accounted for all the variation in the Pb then the sum of the $a_{Pb_j}^2 = 1$ (j = 1 to 5). In the case of lead, $\sum a_{Pbj}^2$ only achieves a value of 0.67; the remainder is $\mu_{Pb}^2$, suggesting other sources or other atmospheric processes affect the Pb concentrations. For Ni, the sum of the $a_{Ni_j}^2$ was 0.91, indicating that the sources of nickel have, for the most part, been identified. The final form for equation 5 would be:

$$Pb = -0.05F_1 + 0.192F_2 + 0.184F_3 + 0.719F_4 + 0.242F_5 + 0.593\ Y_{Pb}$$

The factor equations are defined by the columns in Table 2.2, and should be examined independently of the results sought from Eq. 5. For example, Factor 1 is defined by the following:

$$F_1 = -0.050\ Pb + 0.107\ Mn + 0.118\ Cu + 0.802\ V - 0.029\ Cd + 0.050\ Zn + 0.274\ Fe + 0.879\ Ni + 0.598\ IPSO_4 \qquad \text{Eq. 7.}$$

Geometrically, Factor 1 and Factor 2 are illustrated by Figure 2.3. It can be seen that the new factors are fairly well defined by two groupings of the tracers indicating that a more interpretable form of the tracers has been identified, and a structure of the data can be explained by fewer dimensions. The ancillary data requirements necessary for interpretation of the results are discussed later in this chapter.

The word rotation should be mentioned here since it is commonly used to simplify the structure of any factor analytical solution. A rotation is the simultaneous movement of the groupings of data around the origin of two factor axes. For the simple example in Figure 2.1, there were Factors 1 and 2. In the case of these data, an orthogonal rotation of approximately 45 degrees was necessary to achieve the axial alignments shown. If properly used, a rotation can quite frequently make the sources represented by the factors more readily intepretable. Going back to our hypothetical example in Figure 2.1, the factors presented were the rotated solution. In Figure 2.4, their unrotated form is shown and it is clear that the coordinate system of the original solution was not in its simplest form, although there is approximately $90^{\circ}$ separating the clusters. Rotating the coordinates about $45^{\circ}$ with a standard mathematical procedure yielded new factor axes that were perpendicular or orthogonal, but were more easily identified. This result was obtained using what is termed in the computer program as a varimax rotation. Rotations are of tremendous value since, if the factors are independent of each other, the attainment of a simple structure yields well defined and interpretable solutions.

**Figure 2.3**

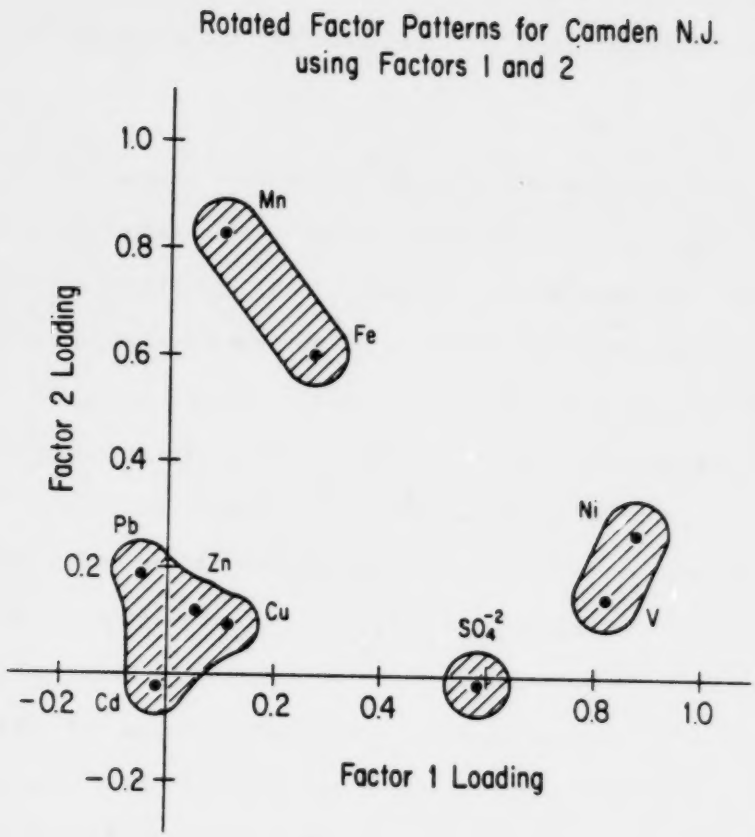
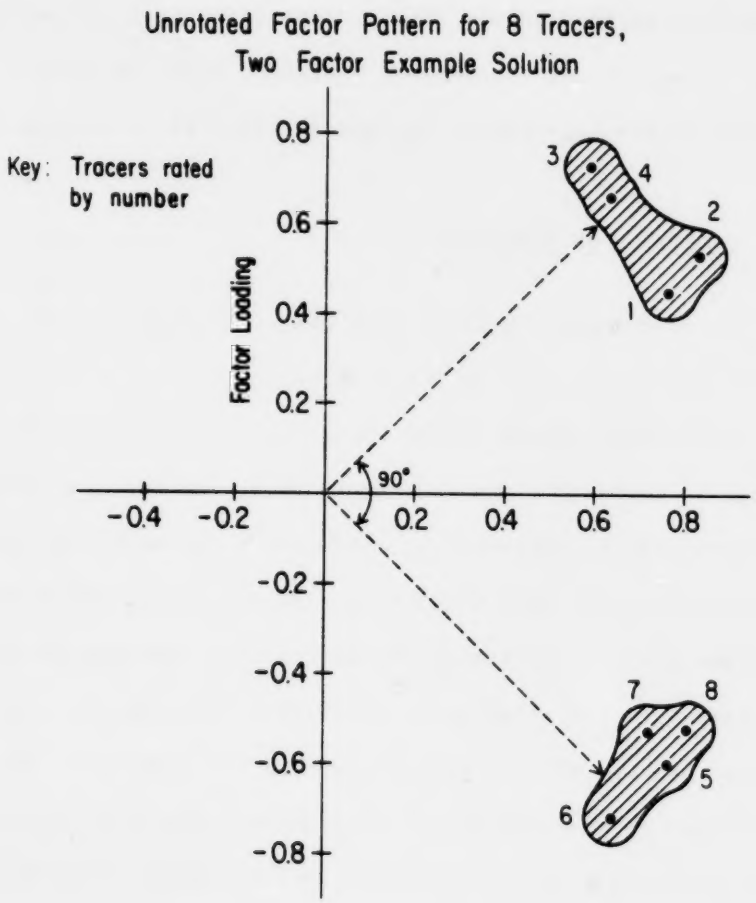

Rotated Factor Patterns for Camden N.J.
using Factors I and 2
Factor 2 Loading
Factor 1 Loading
1.0
0.8
0.6
0.4
0.2
-0.2
-0.2
0.2
0.4
0.8
1.0
Mn
Fe
Pb
Zn
Cu
Cd
Ni
V
$SO_4^{-2}$

Figure 2.4


Unrotated Factor Pattern for 8 Tracers,
Two Factor Example Solution
Key: Tracers rated by number
Factor Loading
0.8
0.6
0.4
0.2
-0.2
-0.4
-0.6
-0.8
-0.4
-0.2
0.2
0.4
0.6
0.8
90°
1
2
3
4
5
6
7
8

Before leaving this topic, it should be stated that an infinite number of rotations are possible which may or may not have perpendicular axes. In some cases, an "oblique rotation" can yield more interpretable solutions when the factors still remain correlated, i.e., the factors or sources are not completely independent of each other.

The above discussion has presented the essentials of the the factor analytical technique. Some of the more important concepts will be described further in the form of applications to air pollution problems and selection of variables. Others will be left to the interested reader in anticipation of his/her application to particular problems.

## 2.2 SELECTION OF VARIABLES

Once the above mathematical basis is conceptually understood, the next important step in the application of factor analytical techniques is the identification of the input tracers. It would be rather naive to suggest that this is a simple task. In fact, if done properly, the investigator is required to have both qualitative and quantitative information on the meteorology, source types, and traffic patterns at his/her disposal to assist in interpreting the derived factors (5). This is essential for allowing an individual to judge the completeness of the tracer set and size of the sample data set needed to identify sources. It cannot be over emphasized that tracer selection is necessary in order for the results to be applicable to the overall questions being asked by the regulatory body.

Major questions that should be asked include:

1) Can the potential major and some minor source be represented by the input variables?

If you are missing trace elements or other marker species that would be related to sources such as soil, oil burning, etc., there is a possibility of ignoring an important source that affects the variance in the ambient data set.

2) Are a sufficient number of cases (sampling periods) available to complete a factor analysis for the tracers used?

A rigorous solution is only achieved when there are greater than a minimum number of degrees of freedom (d) available in the data set. Previous work has been done that indicates the number of cases (n) and variables (v) necessary to satisfy boundary conditions (limits) for the analysis (7). Recently, Henry (20) has suggested that the formula:

$$d/v = n - [(v+3)/2] \quad >30 \qquad \text{Eq. 8}$$

is needed as a minimum precondition in order to attempt a factor model. However, other considerations within this heading include the distribution and transformation of data for each tracer. This would determine if further samples are required to have sufficient data above a detection limit or minimize the effects of extreme values (21).

3) Are there a sufficient number of distinct tracers for different sources within the list available to avoid biasing the solution to a given type of source or a single point source?

The issue of balance among the number of variables was previously mentioned as being important since the dominance of elements

or species associated with a particular source type could bias the solution of the factor model. Since the factor solution is based upon the input data in the correlation matrix, those variables correlated with one source will produce a factor which accounts for a large part of the variation in the chosen set of tracers. Similarly, inclusion of extra tracers related to previously unsuspected sources can yield the identity of a particular source if there is a sufficient number of cases available to decipher its variability. However, an individual source type cannot be identified without a tracer. Henry (20) has suggested that, in practice, multivariate techniques can usually only identify 5 to 8 sources. Therefore, inclusion of large numbers of tracers (> 20) may be fruitless when attempting to interpret the results unless there is a significant increase in the number of samples available for use in the Factor Analyses (see Eq. 8).

4) Are there meteorological data available for the area surrounding the receptor site?

In the initial discussion, it was indicated that factor analytical models apportion the variability of a set of data and meteorology can have a significant effect on those results (5). Therefore, the availability of meteorological components for inclusion in initial attempts at model development may yield improved resolution of individual sources or a major area source. With enough data sets (> 100), it may be possible for the investigator to complete a factor analysis that is stratified by wind direction for individual wind sectors or by heating and cooling degree days.

5) Are the facilities and resources available to conduct further analyses on archived portions of the samples or the analysis of other samples?

Supplemental sample analyses can increase the number of variables by adding elemental or species tracers which could be associated with other source types. In addition, it could provide a means of resolving the source which emit some of the same tracers. Of course, the stability of the samples being considered for supplemental analyses, and the appropriateness of the technique to the type of filter used for sample collection must be fully considered.

The variables of interest for inclusion in source receptor modeling studies include a number of inorganic elements. In addition, Daisey, et al. (22) have suggested that particulate organic species may be used in the future. Similarly, microscopic analyses can yield information on other properties of the sampled particles which can be used to identify sources (4). The techniques available for determining potential tracer elements and species are summarized in many papers on the topic (5,22,24,25). Examples of some source tracers that should be included in a factor analysis variable list for particulate matter are shown in Table 2.3.

The reader is referred to volume V (5) for a list of references on source emission profiles for sources shown in Table 2.3 as well as a number of other source types.

Table 2.3. Sources of Particulate Matter and Tracers.

| Tracers | Potential Source Type |
|---|---|
| V, Ni | Oil burning |
| Al, Si, Ca, Mn, Ti, Fe | Crustal (soil) |
| $SO_4^{-2}$ | Secondary particles |
| Zn, Pb, Cd, As | Primary and Secondary smelting |
| Pb*, Br | Motor vehicles |
| $C^{14}$, K | Wood burning |
| Na, Cl | Marine (ocean) |
| As, Se, Sb | Coal |

*Note: Pb is being phased out as an additive in automobiles which means that a new tracer must be found. Bromine is a scavenger for lead and will also be eliminated as the lead in gasoline decreases.

## 2.3 COMMONLY AVAILABLE STATISTICAL SOFTWARE

1. BMDP Statistical Software

   Manual Available in 1983 Printing

   University of California Press, Berkeley, 1983

   Factor Analytical Programs Include:

   Principal Component Analysis

   Principal Factor (Classical or Common) Analysis

   Maximum Likelihood Factor Analysis

   Kaiser's Second Generation Little Jiffy

   Rotations Available - Seven Rotations Listed Including Varimax and Oblique

2. SAS Software

Manual available: SAS User's Guide, Statistics: 1982 Edition, SAS Institute, Box 8000, CARY, North Carolina 27511

Factor Analytical Programs Include:

Principal Component Analysis

Principal Factor Analysis

Alpha

Maximum Liklihood

Image

Harris Component Analysis

Unweighted Least Squares Factor Analysis

Rotations Available

Seven Options Listed Including Varimax and Oblique

Special Options:

Communalities greater than 1 set to 1 and iteration proceeds

Target pattern rotation

3. SPSS Statistical Software

Manual Available, Statistical Package for Social Sciences, 2nd Ed., McGraw Hill Book Company, New York, 1975, Supplements have been published

Factor Analytical Programs Include

Principal Component Analysis

Principal Factor Analysis

Alpha

Image

RAO

Rotations Available -

Four Options Listed Including Varimax and Oblique

---

For the above statistical packages the following default values are used:

| | SPSS | SAS | BMDP |
|---|---|---|---|
| Number of iterations for initial (unrotated) factor extraction | 25 | 30 | 25 |
| Convergence Criteria for iterations and communality | 0.001 | 0.001 | 0.001 |

These parameters can be respecified by the operator.

The Communalities ($h^2$) for the first three statistical packages are estimated or defaulted by a number of options.

| Package | Communality Estimates ( $h_j^2$) |
|---|---|
| BMDP | |
| Principal Component Analysis | = 1 |
| Principal Factor Analysis | Squared Multiple Correlation (SMC) |
| | Maximum Row Absolute Values (MRAV) |
| | User Specified (US) |
| SAS | |
| Principal Component Analysis | 1 |
| Principal Factor Analysis | SMC; MRAV; US, Random |
| SPSS | |
| Principal Component Analysis | 1 |
| Principal Component Analysis (modified) | SMC; MRAV; US; |
| Principal Factor Analysis | SMC |

4. IMSL Statistical Libraries, Inc., Instructional Manual, 1979 Edition, 7500 Bellaire Boulevard, Houston, Texas 77036

Factor Analytical Programs Include:

Principal Component Analysis

Principal Factor Analysis

Rotations Available

Varimax and Oblique

5. Microcomputer Software -

Programs developed by individual users

Programs by Software Corporations

These vary in options and rotation schemes. However, programs have been written for Principal Component and Factor Analysis with iterative capabilities and rotation available for Apple, IBM, and others. BMDP, SAS and SPSS is also available for specific microcomputer systems.

## 2.4 APPLICATION OF FACTOR ANALYSIS TO AIR POLLUTION DATA AND INTERPRETATION OF RESULTS

### 2.4.1 General Considerations

In developing a Factor Analysis/Multiple regression (FA/MR) Source Apportionment model for a given location, common factor analysis is first used as an exploratory tool to identify sources, to examine relationships between tracer elements or chemical species, and to select statistically independent source emissions tracers. A typical set of air pollution data for a FA/MR model will consist of > 50 measurements of many source tracer species made at a single site over several seasons. Trace elements are the source tracer species most frequently used, but other source tracers such as organic species (22) or microscopic measurements can also be used (4). Ideally, the source tracer species to be used are pre-selected on the basis of what is known about the types of sources in a given area. Existing emissions inventories, the knowledge of local air pollution officials, and microinventories (surveys of the area around a given site) (17) should all be used to identify probable source types affecting the air quality in a given location. Table 2.3 presented an example of some types of sources and typical tracers used to identify those source types. Ideally, measurements of tracers in both the coarse (particles > 2.5 μm in diameter) and fine (< 2.5 μm in diameter) mode of the aerosol should be made as such measurements can be helpful in discriminating sources. For example, resuspended soil and its tracers are found largely in the coarse mode particles while tracers of combustion aerosol are found in fine mode particles. As stated previously, measurement of more than one tracer

for a given source type is also advisable and can be useful in resolving sources in those instances in which a preselected tracer is found to originate from other source types than expected. For example, Pb might originate from industrial as well as motor vehicle sources; by including measurements of the second tracer, Br, it is usually possible to identify the presence of both source types.

In many instances, an existing set of trace elemental data can be used to identify major sources impacting a given site. In this case, it is advisable to select subsets of variables which are appropriate tracers for source types of interest, with an emphasis on having balance in the number of tracers per source.

### 2.4.2 Preliminary Examination of the Data

Prior to attempting a factor analysis, the data should be carefully reviewed. Limits of detection, uncertainties in measurements, and extreme outliers in the data set should be examined. Measured variables close to the limits of detection of a particular measurement technique will have greater variability due to the analytical uncertainties than due to variations of elemental composition in the sources. Consequently, data of this nature will be of little use in the model development. Extreme outliers in a data set often indicate errors in data entry or unusual events, such as episodes. Since the correlation coefficients, which are used as the first step in the development of a factor analytical model, can be strongly biased by a single extreme value, it is advisable to examine the distribution of the measurements of each variable and to eliminate such outliers from the data set prior to factor analysis. Figure 2.5 presents some examples of the distributions of

ambient measurements for several tracers. The distribution of the Ni (2.5.1) shows that a substantial fraction of the data are below or near the analytical limit of determination (LLD). The tail of the distribution, to the right of Figure 2.5.1, indicates that there are also a large number of high values; this distribution is obviously far from normal. The distribution of Mo values in Figure 2.5.2. appears more normally distributed; however, the midpoint of the distributions marks the analytical limit of determination for Mo. In this instance, the uncertainties in the analytical determinations can be expected to account for most of the variation in the data.

Figure 2.5.3 presents a distribution of measured Fe data which more closely approximates a normal distribution. The LLD for Fe was 30 $ng/m^3$ and all of the measurements were above this value. The reason for the extremely high concentrations should be investigated separately since they may identify a strong point source or unusual meteorological event. The examples given here represent only part of a print out from a BMDP program which can be used to examine the distributions of measurements prior to attempting a factor analysis. In the preliminary examination of data, the entire output of the distribution program would be carefully examined.

### 2.4.3 Factor Analysis Solutions

For the FA/MR model, classical factor analysis with rotation, in most cases Varimax, is used to identify sources and source tracers. Common factor analysis assumes that the variations in a given data set

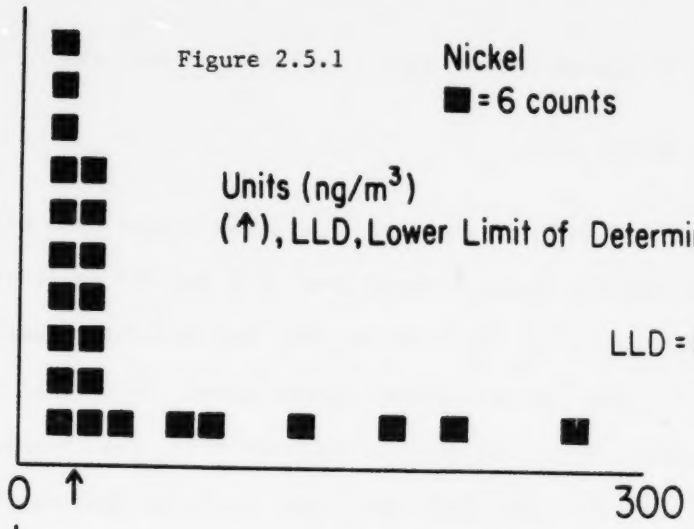



Figure 2.5.1

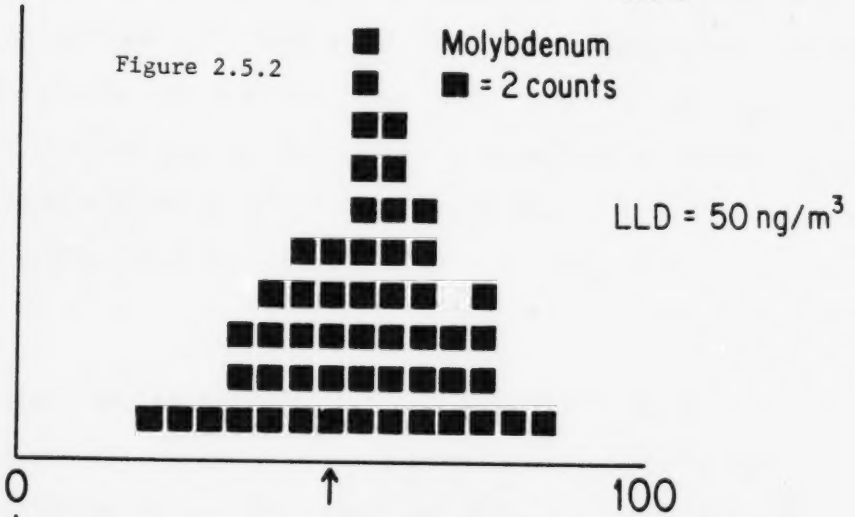



Figure 2.5.2

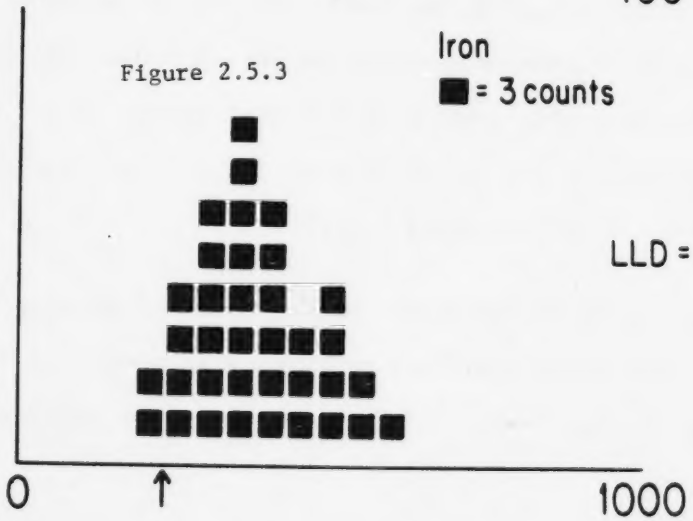



Figure 2.5.3

are of two types:

a) common or shared determinants, such as sources; and

b) unique determinants.

The unique part of the variation of a given tracer does not contribute to relationships among tracers available for the analysis and can be due to uncertainties in the analytical measurements, unidentified sources, or changes in emissions composition. Because of common or shared determinants among tracers, it follows that the common determinants will account for the observed relations between the source tracers. As previously indicated, these will be smaller in number, i.e., some tracers will come from the same type of source. Consequently, these can be expressed as a reduced set of variables called the factors or common determinants. Each factor would be defined in an equation similar to Equation 4 as a mathematical combination of the original tracers.

A key problem in factor analysis is determining how many factors should be retained, i.e., how many are meaningful. The number of factors obtained is determined by the number and type of variables included and by the minimum eigenvalue selected for the analysis. The eigenvalue is a characteristic number associated with each factor; it is also the sum of the squares of the factor loadings ($a_{ij}$) of each variable for a given factor (the columns in Table 2.1).

The default minimum eigenvalue for most computer programs is 1.0; this, however, has often been found to be too high a value to obtain all of the sources of significance. In general, the best approach is to

view the factor analysis as an exploratory technique and, using different criteria, to factor analyze various subsets and combinations of variables and cases.

As a first step, the entire set of variables can be analyzed using the computer program default criteria. These include a minimum eigenvalue, the number of iterations to reach the final estimated communalities and the value for convergence of the communalities (i.e., the communalities are iteratively estimated until the values change by no more than a specified value, and, as shown on page 27, the default convergence criterion is usually 0.001).

The communalities of the variables and eigenvalues for the first set of factors calculated (one per variable) as well as the factor loadings of the rotated factors and factor score coefficients of each case should be examined in the computer print out. If the estimated communality of any variable has exceeded 1.0, the iterative procedure of the program will be terminated before the convergence criterion is reached and the rotated factors obtained may be misleading. Alternatively, the default value for number of iterations may not be high enough to reach the convergence point. The default eigenvalue of 1.0 may be too high to identify all factors of importance and thereby underestimate the rank of the correlation matrix. Subsequently, it may be appropriate to select a lower value for the next factor analysis. An eigenvalue of 0.7 or 0.8 or even as low as 0.5 may then be selected to determine the number of factors to be retained.

### 2.4.4 Types of Factors

The factors and factor loadings must also be carefully examined to identify the type of factor obtained. There are a number of types of factors which are commonly obtained with air pollution data, many of which have been discussed by Roscoe, et al (26) for principal component factor analysis. In general, the discussion is applicable to common factor analysis.

1. Single source-type factor (area source) - due to emissions from a single type of source, such as motor vehicles, and readily identified by high factor loadings (0.70 - 0.99) of characteristic tracer elements.

2. Coincident multiple source factor - due to two or more sources or source types having similar temporal variations in emissions patterns, having similar impacts on the receptor site due to their locations and prevailing meteorology or having similar emissions compositions. Factor loadings tend to be somewhat lower for this type of factor and source tracers are mixed, i.e., the tracers for both source types are found on a single factor.

3. Anti-coincident multiple source factor - usually due to sources requiring mutually exclusive meteorological conditions to impact on the receptor site. These are characterized by high negative as well as positive loadings for certain elements on the same factor.

4. Single source factor (point source) - due to emissions from a single source with a unique tracer(s). Such a factor usually accounts for a small percentage of the total variance in the data and can be recognized by factor scores which are generally low, but with a number of very high scores for certain days or samples. Morandi, et al (17) have observed such a phenomenon for a zinc smelter in Newark and noted that the concentrations of zinc were unusually high on certain days when the prevailing winds were from the northeast.

5. Regional Transported Aerosol Factor - due to aerosol transported from an area upwind of the site. This type of factor has been reported by Thurston (16) and was characterized by the high loadings for the tracer elements Se, S and Mn and the high correlation between the factor loadings for this factor and certain types of meteorological conditions associated with the transport of sulfate aerosol into the northeastern U.S.

6. Error Factors - due to random errors, individual data point errors, bias errors or errors in the use of factor analysis. Random error and individual data point errors can usually be identified and prevented through the preliminary data screening. Bias errors can be due to the analysis method or may occur due to sampling artifacts, e.g., volatilization, condensation or chemical change during sampling. Errors in the use of factor analysis can be more difficult to identify but generally occur as a consequence of inexperience with the technique.

## 2.5 AN EXAMPLE OF THE APPLICATION OF FACTOR ANALYSIS TO AIR POLLUTION DATA.

Tables 2.4 and 2.5 present the results of a factor analysis on a set of data as an example. This data set was not subjected to the recommended preliminary data examination, but the results have been included to point out certain features and pitfalls of factor analysis. In this instance, an eigenvalue of 0.7 was used to determine the number of retained factors. Inspection of Table 2.4 reveals that more than 25 iterations are required for convergence of the estimated communalities to 0.001. It also reveals that the factors beyond factor 6 each account for less than 5% of the total variance in the data set. The high factor loadings of the rotated factor matrix, Table 2.5, help identify the following types of sources: Factor 1 - Pb, Br; motor vehicles; Factor 2 - Fe, Ti; soil-related; Factor 3 - V, P - oil burning; Factor 4 - S; sulfate aerosol,. Factor 5, with high loadings for V and Ni also appears to be related to oil burning while Factor 6 has a high loading only for molybdenum. These last two factors account for 6.3% and 3.3%, respectively, of the total variance. The communalities for both Mo and Ni are less than 0.5. Examination of the analytical limits of determination for these variables and their distributions reveals that most of the Mo measurements and about 20% of the Ni measurements are less than or close to the analytical limits of determination. In addition, three extremely high values of Ni (> 100 ng/m$^3$) were in the data set. Thus, it would appear that factors 5 and 6 are of the type related to errors in the data, as discussed by Roscoe, et al (26).

---

**The analytical limit of determination is that value below which the concentration of a species cannot be reliably determined.

Table 2.4. Results of Initial Factoring.

| Variable | Estimate of Communality | Factor | Eigenvalue | Percent of Variance |
|---|---|---|---|---|
| $SO_4^=$ | 0.511 | 1 | 2.18 | 24.2 |
| Fe | .582 | 2 | 1.94 | 21.5 |
| Pb | .806 | 3 | 1.64 | 18.2 |
| Mo | .198 | 4 | 1.06 | 11.7 |
| P | .354 | 5 | 0.79 | 8.8 |
| Br | .826 | 6 | 0.70 | 7.8 |
| V | .557 | 7 | 0.36 | 4.0 |
| Ti | .571 | 8 | 0.24 | 2.7 |
| Ni | .242 | 9 | 0.08 | 0.9 |

More than 25 iterations required

Table 2.5. Varimax Rotated Factor Matrix[a]

| Variable | Communality | Factors and Factor Loadings 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| S | 0.632 | -.014 | 0.106 | 0.126 | 0.747 | 0.099 | 0.194 |
| Fe | 0.814 | 0.114 | 0.874 | 0.099 | 0.055 | 0.001 | 0.159 |
| Pb | 0.980 | 0.951 | 0.054 | -0.102 | 0.247 | 0.019 | -0.025 |
| Mo | 0.451 | -0.005 | 0.207 | -0.064 | 0.179 | -0.042 | 0.608 |
| P | 0.592 | -0.060 | 0.187 | 0.686 | 0.164 | 0.157 | -0.180 |
| Br | 0.969 | 0.919 | -0.106 | 0.103 | -0.315 | 0.052 | 0.028 |
| V | 0.877 | 0.119 | -0.023 | 0.677 | -0.034 | 0.588 | 0.239 |
| Ti | 0.692 | -0.147 | 0.800 | 0.072 | 0.093 | -0.072 | 0.103 |
| Ni | 0.414 | 0.016 | -0.045 | 0.148 | 0.084 | 0.615 | -0.075 |
| Probable Source | - | Motor Vehicles | Soil | Oil Burning/ Phosphorous | Sulfate Aerosol | Oil Burning | Molybdenum |

[a]n = 93 cases.

The relatively high negative loading of Br (-0.315) on factor 4 should also be noted. This is suggestive of a sampling artifact due to volatilization losses of Br (probably as HBr) in the presence of $H_2SO_4$ aerosol (27).

Having examined the results shown in Tables 2.4 and 2.5, several other factor analyses were performed on different subsets of variables, with Mo and Ni omitted, and sets of cases (sampling days) which had more normally distributed data (all data were within three standard deviations of the mean). An example of these is presented in Table 2.6. Five factors were obtained in this instance and are clearly identifiable as being related to resuspended soil (Fe, Ti), motor vehicles (high factor loadings for Pb and Br), sulfate-related aerosol ($SO_4^=$), oil burning (V), and incineration (Cu). These results were consistent with what was known about the major sources of airborne particles at the site where the samples were collected for analysis. The tracers, Fe, Pb, S, V and Cu have high loadings on each of the factors, and the factors are orthogonal (statistically of one another). These independent tracers were then selected as predictors for the next step in the development of the FA/MR model, that is, the development of a stepwise multiple regression model. The results of this exercise are shown in Section 3.3. The quantitative regression model will then be used to estimate the contributions of each source type to the ambient aerosol concentration.

Table 2.6. Factor Loadings of Rotated Factor Matrix

| Variable | Communality | Factors 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| $SO_4^=$ | 0.921 | -0.024 | -0.006 | 0.951 | -0.089 | 0.090 |
| Fe | 0.917 | 0.949 | 0.068 | 0.011 | 0.101 | 0.038 |
| Pb | 0.767 | -0.042 | 0.852 | 0.113 | -0.001 | 0.165 |
| Br | 0.724 | -0.090 | 0.832 | -0.139 | 0.063 | -0.007 |
| V | 0.549 | -0.058 | 0.113 | -0.360 | 0.614 | -0.162 |
| Ti | 0.684 | 0.792 | -0.216 | -0.034 | 0.070 | 0.065 |
| Mn | 0.647 | 0.437 | -0.042 | 0.138 | 0.653 | 0.094 |
| Cu | 0.366 | 0.055 | 0.085 | 0.066 | -0.032 | 0.592 |
| % of Variance | -- | 35.3 | 26.6 | 22.3 | 10.3 | 5.6 |
| Probable Source Type | | Resuspended Soil | Motor Vehicles | Sulfate | Oil Burning | Incineration |

a. n=78 cases; eigen value cut off = 0.8; convergence required 51 iterations.

## 2.6 VALIDATION OF FACTOR ANALYSIS

### 2.6.1 Introduction

The models developed by the existing methods for source apportionment require validation if any meaningful use is to be made of the results. Some models permit relatively simple comparisons to actual data. Measurement of the growth of a tree, enumeration of organisms in a lake *vs.* time, or measurement of crop yields may all serve to confirm model outputs.

In receptor modeling, the alternative means of evaluating the truth of a model's output are generally the results of other models, the internal consistency of the model and the consistency of the model with what is known about the severity of the air pollution problem in a given area. Thus, some care is required when attempting to validate the model.

Various estimates can be made for the expected concentration of a pollutant. Emission inventories can indicate the importance of each source or source type evaluated, and dispersion modeling of the inventory emissions provides estimates of the pollutant concentrations. When such estimates are compared to measured data, the accuracy of the inventory and dispersion modeling also can be evaluated. In some cases increasing divergences have occurred between dispersion model estimates and measured TSP values as TSP levels declined because of effective pollution control measures. Efforts to add non-point sources and improve the dispersion model have been partially successful. However, source apportionment by the multivariate statistical approach offers an

attractive and potentially less expensive alternative and can identify contributing source types without the need for extensive source testing in a potential impact area.

The various source apportionment-receptor models, such as the FA/MR or chemical mass balance (CMB) models (2), may be intercompared where data sets support the use of both mathematical approaches. The regression models from FA/MR analysis yield estimates of source contributions and regression coefficients relating the source tracer to the emitted mass. These coefficients are subject to verification.

. Finally, all models are subject to re-evaluation where changes in source types or quantities of emissions are observed or predicted. Model results, if accurate, can be used to predict trends. This has been successful in the past for simple emission inventory estimates. Where trends in concentrations of the predictor tracers, the TSP or any other particulate mass measure, are observed the model derived from any source apportionment method should accurately predict or reflect the observed changes.

### 2.6.2 Source Composition Profiles

The composition of particles emitted by a given source can be obtained by analysis of a sample of the stack emissions collected on a filter in a standard EPA train (28). More recently it has been noted that several toxic elements and many organic compounds are condensed onto the particles as the stack gases cool (29). This may continue to occur for some distance in the plume after leaving the stack (30).

Composition profiles useful at receptor sampling locations are likely to be somewhat different from the emissions data obtained by sampling heated stack gases. As plume sampling is very difficult and very expensive, few useful plume samples have been obtained. Emission composition profiles previously cataloged are possibly biased and of somewhat doubtful use. Development of a suitable set of profiles in a complex urban-industrial area would be very difficult and extremely expensive. Thus the applications of factor analysis methods to ambient sample composition data has provided an alternative means of obtaining data which relates ambient compositions to the probable sources of the emissions.

### 2.6.3 Factor Validation

Factor analysis has been used as an exploratory tool to select possible source tracers for New York City (9, 31) and for Boston (32). In each of these early applications of the method, comparisons were made among different source emission profiles to confirm that the factor was associated with one source type. Kleinman *et al.* (9) found factors with elemental loadings which indicated automobiles, oil burning, soils, incinerators and a source of a sulfate compound were the types affecting New York City. However, it should be noted that the factors did not reproduce the source profile, since z-scores were used in the factor analysis.

Emission sample composition analysis gives a source composition profile with a concentration for each element in mass per unit mass of emitted particles. Derived factors provide loadings or scores which indicate the degree to which the variability (variance) of the tracer element is related to the factor. The most strongly loaded element may

provide an independent predictor tracer variable, but may not be a major mass contributor in the source emissions. For instance, selenium may be a useful tracer for coal fly ash, while the concentrations of major elements Al, Si, Fe, Ca, Mg, etc. are so similar to those in soils that they are useless as predictor variables. Moreover, any trace elements which are loaded at equal levels on several factors are unlikely to be useful as independent predictor variables.

From existing emission profiles, certain elements are expected to load significantly on separate factors. For instance, Pb, V, and $SO_4^=$ are typically assigned to autos, oil burning and secondary aerosol sources. Lead may be emitted by primary and secondary smelters yielding factors with Pb-Br for autos and some combination of Pb-Zn-Cu-As loadings when smelters are involved. Should steel industries be present, a strongly loaded Fe-Mn factor may occur. However, Fe and Mn should also load with Si, Al, Ca, Ti on a soil factor unless the industrial contributions are overwhelming. Should more complex situations occur, more exact comparisons of emission and ambient profiles may be made by conversion of the factor scores to relative concentration profiles as has been done in the methods described by Thurston (33), Dattner (in Currie et al., 34) and Hopke et al. (32, 35).

### 2.6.4 Evaluation of Source Inventories

Where the factors are not readily related to previously reported source profiles, a careful macro and micro source inspection may be in order. This serves both to identify sources or source types previously omitted from past studies and to confirm the presence of sources identified by the factor analysis approach but not listed in the category of

expected sources for the sampling site.

In a study of Newark, N.J., Morandi et al. (12) observed four to seven factors (Table 2.7), depending upon the variables included in the analysis. The table is provided for illustrative purposes, and a detailed discussion is given in the reference. Factors and predictor tracers were easily identified for oil burners, soil, industrial sources, the sulfate related mass and autos were identified in the final analysis - Row 6. Two factors were found with unusual loading patterns. One was loaded with zinc and benzo(e)pyrene and a second contained only a seasonal component. The regression model coefficient for zinc indicated the composition to be ZnO, a likely emission from a smelter. This was verified from state emissions data which showed that a single zinc smelter existed nearby and was in operation.

### 2.6.5 Factor Stability

Emission tracers established by the factor loadings may match with published source sample data for carefully selected particle size ranges stable, nonvolatile elements or substances. Factor analysis selections of independent predictor tracers may differ from predictions based on source profiles, where source and ambient particle size ranges are not matched, or volatile, condensible and reactive substances are involved. In these cases the variations may be examined by compiling "source" profiles from the factors with the methods used by Dattner (34) or Thurston (33). Calculation of the ratio of a potentially variable substance to a stable element or substance normalizes for dilution or other effects which do not cause relative profile shifts. Equivalent ratios for source and ambient data indicate no change in relative composition.

Table 2.7 Factor Analysis of the Newark, N.J. Data Set.

| Variables Entered | N | F1 | F2 | F3 | F4 | F5 | F6 | F7 |
|---|---|---|---|---|---|---|---|---|
| Pb, Mn, Cu, V, Cd, Zn, Fe, Ni, $SO_4^{-2}$ | 77 | V(0.94) Ni(0.92) Pb(0.47) Mn(0.46) | Cu(0.76) Pb(0.75) | Cd(0.73) Fe(0.52) | Fe(0.56) Mn(0.54) | | | |
| Probable Source | | oil burning/ space heating | motor vehicles | unknown | soil | | | |
| Pb, Mn, Cu, V, Cd, Zn, Fe, Ni, $SO_4^{-2}$ COmax | 77 | Mn(0.90) Fe(0.77) | V(0.83) Ni(0.82) | Pb(0.79) COmax(0.69) | Cu(0.77) Fe(0.51) | $SO_4^{-2}$(0.52) | Zn(0.54) | |
| Probable Source | | Soil | oil burning/ space heating | motor vehicles | smelter/metal related indus. | $SO_4^{-2}$/secondary aerosol | zinc related source | |
| Pb, Mn, Cu, V, Cd Zn, Fe, Ni, $SO_4^{-2}$ COmax, $SO_2$Ave | 66 | V(0.83) Ni(0.82) | COmax(0.78) Pb(0.66) $SO_2$Ave(0.59) | Mn(0.83) Fe)0.67) | Cu(0.81) Fe(0.57) Pb(0.42) | $SO_4^{-2}$(0.50) | Zn(0.50) $SO_2$Ave(0.43) | |
| probable Source | | oil burning/ space heating | motor vehicles | soil | smelter/metal related indus. | $SO_4^{-2}$/secondary aerosol | zinc related source | |
| Pb, Mn, Cu, V, Cd, Zn, Fe, Ni, $SO_4^{-2}$ IBEP, IBAP, COmax | 57 | Fe(0.93) Mn(0.83) Cu(0.75) Cd(0.63) | V(0.87) Ni(0.87) | COmax(0.71) Pb(0.68) BAP(0.64) | BEP(0.87) Zn(0.83) | $SO_4^{-2}$(0.54) | | |
| Probable Source | | soil | oil burning/ space heating | motor vehicles | zinc related source | $SO_4^{-2}$/secondary aerosol | | |

COmax = maximum hourly carbon monoxide concentration.
$SO_2$Ave= 24 hour average $SO_2$ concentration.
BAP = Benzo(a)pyrene.
BEP = Benzo(e)pyrene.
Headd = heating degree days.
Coodd = cooling degree days.

Based on the fact that larger particles have higher setting velocities, a ratio of the airborne concentrations of an element concentrated in large particles to that of a second element concentrated in small particles will decrease with time of residence in the atmosphere. This assumes no fresh infusion of new particles. Volatile materials may condense as a plume cools or shift from particle to vapor phase and back again as ambient temperatures increase and decrease. Reactive materials can decrease in concentration with time.

Careful examination of composition profiles and ratios of nonstable to stable tracers should show shifts between source emission test profiles and ambient profiles which compare well to expected deposition rates and volatility or reaction characteristics. Shifts which do not agree qualitatively are indicative of possible difficulties in selection of independent predictor variables by the factor analysis method.

## 2.7 References

1. U.S. EPA, 1981a: Receptor Model Technical Series Volume I: Overview of Receptor Model Application to Particulate Source Apportionment. EPA-450/4-81-016a.

2. U.S. EPA, 1981b: Receptor Model Technical Series Volume II: Chemical Mass Balance. EPA-450/4-81-016b.

3. U.S. EPA, 1983a: Receptor Model Technical Series Volume III: User's Manual for Chemical Mass Balance Model. EPA-450/4-83-014.

4. U.S. EPA, 1983b: Receptor Model Technical Series Volume IV: Summary of Particle Identification Techniques. EPA-450/4-83-018.

5. U.S. EPA, 1984: Receptor Model Technical Series Volume V: Methods for Combining the Various Source Apportionment Approaches (in press).

6. Wolff, G.T., 1980, Mesoscale and Synoptic Scale Transport of Aerosols. Ann. N.Y. Acad. Sci., T.J. Kneip and P.J. Lioy, Eds., 338: 379-388.

7. Harman, H.H., 1976, Modern Factor Analysis, 3rd Ed., University of Chicago Press, 1-487.

8. Rummel, R.J., 1970, Applied Factor Analysis, Northwestern University Press.

9. Kleinman, M.T., B.S. Pasternack, M. Eisenbud, and T.J. Kneip, 1980, Identifying and estimating the relative importance of sources of airborne particulates. Environ. Sci. Technol., 14: 62-65.

10. Dzubay, T.G., R.K. Stevens, W.D. Balfour, H.J. Williamson, J.A. Cooper, J.E. Core, R.T. Decesar, E.R. Crutcher, S.L. Dattner, B.L. Davis, S.L. Heisler, J.J. Shah, P.K. Hopke and D.L. Johnson, 1984, Interlaboratory comparison of receptor model results from Houston aerosol. Atmos. Environ. (in press).

11. Stevens, R.K., T.G. Dzubay, C.W. Lewis and R.W. Shaw, Jr., 1984, Source apportionment methods applied to the determination of the origin of ambient aerosols that affect visibility in forested areas. Atmos. Environ., 18: 261.

12. Morandi, M., J.M. Daisey and P.J. Lioy, 1983, A receptor source apportionment model for inhalable particulate matter in Newark, N.J. Paper No. 83-14.2 presented at the 76th Annual Meeting of the Air Pollution Control Association, Atlanta, GA, June 19-24.

13. Pace, T.G. The Role of Receptor Models for Revised Particulate Matter Standards, IN: Receptor Models Applied to Contemporary Pollution Problems, S.L. Dattner and P.K. Hopke, Eds., Air Pollution Control Association, Pittsburgh, PA, pp. 18-28.

14. Henry, R.C. and G.M. Hidy, 1979, Multivariate analysis of particulate sulfate and other air quality variables by principle components - Part I. Annual data from Los Angeles and New York. Atmos. Environ. 13: 1581-1596.

15. Lioy, P.J., R.P. Mallon, M. Lippmann, T.J. Kneip and P.J. Samson, 1982, Factors affecting the variability of summertime sulfate in a rural area using principal component analysis. J. Air Poll. Contr. Assoc. 32: 943-1047.

16. Thurston, G.D. and J.D. Spengler, 1982, Source contributions to inhalable particulate matter in metropolitan Boston, Massachusetts. Paper No. 82-21.5 presented at the 75th Annual Meeting of the Air Pollution Control Association, New Orleans, Louisiana, June 20-25.

17. Morandi, M.T., 1985, Development of Receptor Oriented Source Apportionment Models for Inhalable Particulate Matter and Particulate Organic Matter in New Jersey. Ph.D. Dissertation, New York University Medical Center, February.

18. Guttman, L., 1954, Some necessary conditions for common factor analysis. Psych. 18, 277-296.

19. P.J. Lioy and J.M. Daisey, 1984, ATEOS Project. Unpublished Data.

20. Henry, R.C., C.W. Lewis, P.K. Hopke and H.J. Williamson, 1984, Review of receptor model fundamentals. Atmos. Environ. (in press).

21. Draper, N. and H. Smith, 1981, Applied Regression Analysis, 2nd Ed., Wiley Interscience, New York, 1-709.

22. Daisey, J.M., 1983, Receptor Source Apportionment Models for Two Polycyclic Aromatic Hydrocarbons. IN: Receptor Models Applied to Contemporary Pollution Problems, S.L. Dattner and P.K. Hopke, Eds., Air Pollution Control Association, Pittsburgh, PA, pp. 348-357.

23. Watson, J.G., 1979, Chemical element balance receptor model methodology for assessing the source of fine and total suspended particulate matter in Portland, Oregon. Ph.D. Dissertation, Oregon Graduate Center, February.

24. Miller, M.S., S.K. Friedlander and G.M. Hidy, 1972, A chemical element balance for the Pasadena aerosol. J. Colloid Interface Sci., 39: 165-176.

25. Gordon, G.E., W.R. Pierson, J.M. Daisey, P.J. Lioy, J.A. Cooper, J.G. Watson, Jr. and G.R. Cass, 1984, Considerations for design of source apportionment studies. Atmos. Environ. (in press).

26. Roscoe, B.A., P.K. Hopke, S.L. Dattner, and J.M. Jenks, 1982. The use of principal component analysis to interpret particulate compositional data sets. J. Air Pollut. Control Assoc., 32: 637-642.

27. Pierson, W.R. and W.W. Brachaczek, 1983. Particulate matter associated with vehicles on the road. II. Aerosol Sci. Technol., 2: 1014.

28. U.S. EPA, 1971, Method 5, Federal Register, 36(247), p. 24880.

29. Natusch, D.F.S., 1978. Potentially carcinogenic species emitted to the atmosphere by fossil-fueled power plants. Environ. Health Pers., 22: 79-90.

30. Natusch, D.F.S. and Tomkins, B.A., 1978. Theoretical consideration of the adsorption of polynuclear aromatic hydrocarbon vapor onto fly ash in a coal-filed power plant. Carcinogenesis, Vol. 3: Polynuclear Aromiatic Hydrocarbons, P.W. Jones and R.I.

Freudenthal, Eds., Raven Press, New York, 145-153.

31. Kneip, T.J., M.T. Kleinman and M. Eisenbud, 1973. Relative contribution of emission sources to the total airborne particulates in New York City. Proc. Third. Int. Clean Air Congress, Dusseldorf.

32. Hopke, P.K., 1980. Source identification and resolution through application of factor and cluster analysis. Ann. N.Y. Acad. Sci., 338: 103-115.

33. Thurston, G.T. and J.D. Spengler, 1985. A quantitative assessment of source contributions to inhalable particulate matter pollution in metropolitan Boston. Atmos. Environ, 19 (in press).

34. Hopke, P.K., D.J. Alpert and B.A. Roscoe, 1983. Fantasia - A program for target transformation factor analysis to apportion sources in environmental samples. Computers and Chemistry, 7: 149-155.

35. Currie, L.A., et al., 1984, Intercomparison of source apportionment procedures: results for simulated data sets. Atmos. Environ. 18: 1517-1537.

## 3.0 MULTIPLE REGRESSION ANALYSIS

After identifying the major sources of particulate pollutants and selecting independent source tracers or predictors using common factor analysis, the next step is to obtain a quantitative relationship between particle mass concentration and the concentrations of the source tracers. Such a relationship will provide estimates of the contributions of each source or source type to atmospheric concentrations of particulate matter and can assist in determining control strategies. This relationship is determined by stepwise multiple regression analysis. Although the mathematical steps in obtaining such a relationship and the application of such analysis to real world data can be quite complex, the conceptual basis is fairly straightforward. In order to acquaint the reader with the fundamental concepts and some of the terms used in multiple regression analysis, an example of a simple (one dependent and one independent variable) bivariate linear regression model is presented and discussed first. In the particulate mass air pollution example, the independent variable would be an elemental tracer and the dependent variable would be the total particulate mass or a fraction of the particulate mass. This is then extended to a simple multiple linear regression model where more than one independent tracer is used. Finally, an example of the actual application of stepwise multiple linear regression to air pollutant data is given using the results from the factor analysis presented in Table 2.6 to select tracers and the source attributions identified.

A more thorough understanding of multiple regression is required for those who actually develop multiple regression models. Standard

references on this technique, such as that by Draper and Smith (1) should be consulted.

## 3.1 SIMPLE BIVARIATE LINEAR REGRESSION MODEL

In simple regression analysis, which is familiar to many in the field of air pollution, the values of a dependent variable, Y, are predicted from a linear relationship to an independent predictor variable X. This relationships is of the form:

$$Y = A + BX \qquad \text{Eq. 1.}$$

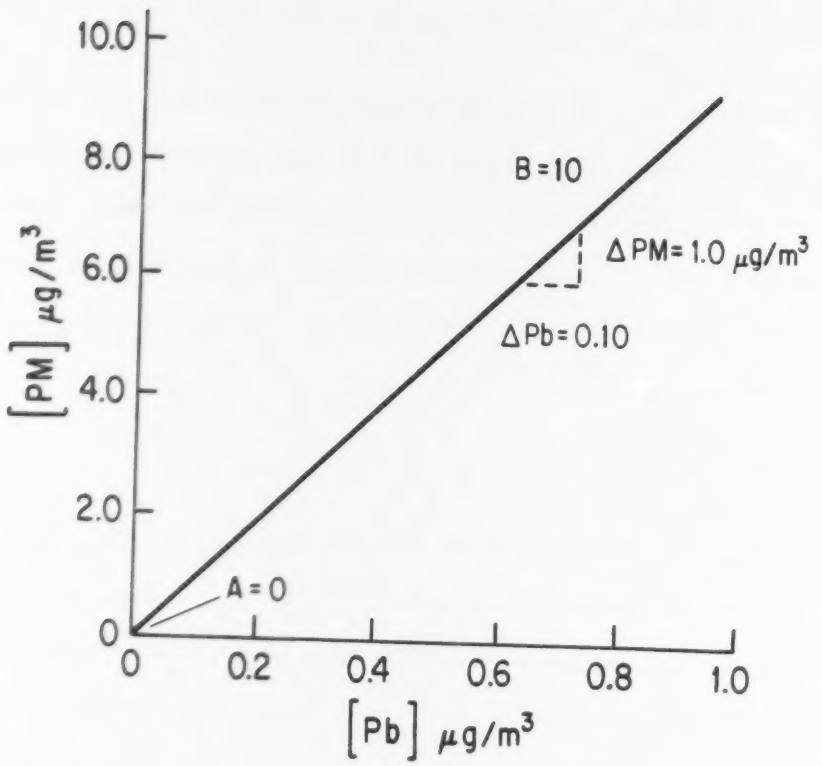
where B is a constant by which all values of X are multipled, and A is a second constant which is added in each case to predict Y.

As an example, assume that for a given site, only motor vehicles using leaded gasoline contribute to airborne concentrations of particle mass less than 10 μm, $PM_{10}$, and that Pb can be used to trace these emissions. Thus, when Pb concentrations increase, concentrations of $PM_{10}$ increase proportionately at this site. The model, in this instance, would be:

$$[PM] = A + B[Pb], \qquad \text{Eq. 2.}$$

where [PM] and [Pb] are the concentrations of $PM_{10}$, and Pb, respectively, B is the slope of the line and A is the intercept. If there were no errors in the measurements of $PM_{10}$ and Pb, and the model perfectly described the sources of $PM_{10}$ for this site, then A would be equal to zero. When the intercept is zero, the result implies that all of the $PM_{10}$ is proportional to Pb, and that there is no other "unexplained source". This is shown graphically in Figure 3.1. The

FIGURE 3.1 Idealized Linear Relationship Between PM-10 and Pb




| PM . . µg/m³ |
|---|
| 2.0 |
| 3.0 |
| 4.0 |
| 6.0 |
| 8.0 |
| 10.0 |

| Pb . µg/m³ |
|---|
| 0.20 |
| 0.30 |
| 0.40 |
| 0.60 |
| 0.80 |
| 1.00 |

coefficient B would then be determined simply by dividing each measured value of IPM by its paired measured Pb value; B would be constant for each case and would describe the ratio of $PM_{10}$ to Pb in motor vehicle emission. The value of B is 10 in this example, i.e., Pb is 0.10 of the particle mass emitted by these ideal motor vehicles.

Since measurements always involve some uncertainties and the real world is far from ideal, the values of B determined would vary somewhat from case to case. The value of A will also differ from zero. If average values of B and of A were selected to predict $PM_{10}$ from the concentration of Pb, the individual predicted values of $PM_{10}$ (PM) would differ from the actual value, i.e.,

$$[PM] - [\hat{PM}] \neq 0 \qquad \text{Eq. 3.}$$

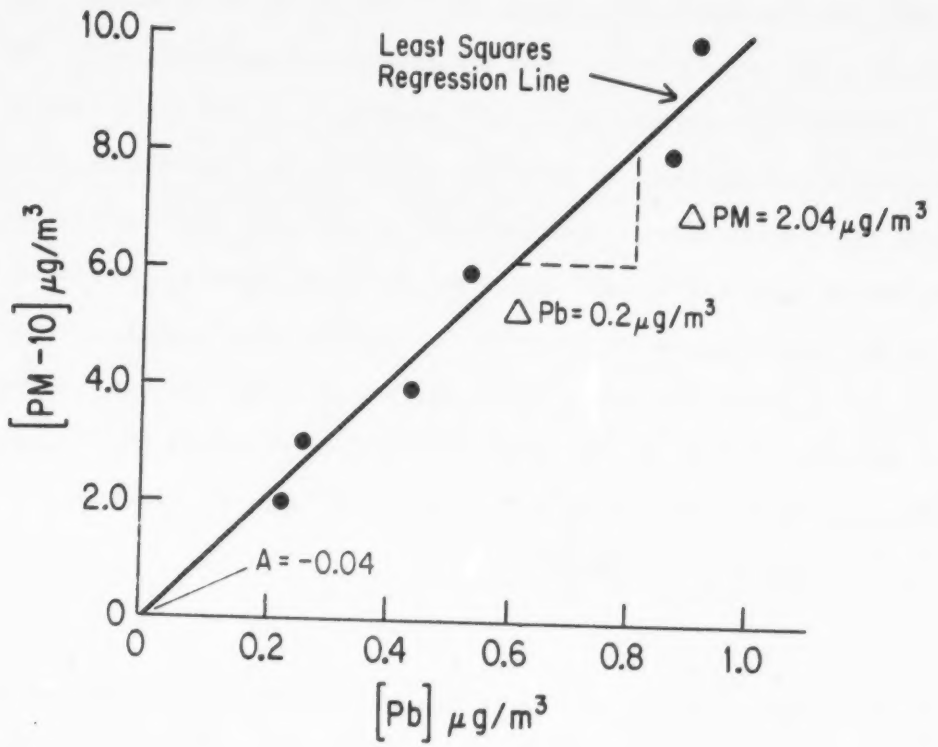
The value of this difference between actual and predicted $PM_{10}$ for each case is termed the residual. Analysis of residuals often give useful information about the goodness of fit of the model, and patterns or correlations which indicate a need for additional source terms. A thorough discussion of Residual Analysis can be found in Draper and Smith, 1981.

Regression analysis involves determining A and B in such a way that the sum of the squares of these residuals is a minimum;

$$(PM - \hat{PM})^2 = \text{minimum} \qquad \text{Eq. 4.}$$

Thus, the values of A and B which define the linear relationship between Pb and $PM_{10}$ are selected to minimize the deviation of the actual values of $PM_{10}$ from the line shown in Figure 3.2 these values can be calculated

**FIGURE 3.2 Linear Relationship Between PM-10 and Pb With Uncertainties in Measured Concentration of PM-10.**




| PM, μg/m³ Actual | PM, μg/m³ Predicted[a] | Pb, μg/m³ |
|---|---|---|
| 2.0 | 2.21 | 0.220 |
| 3.0 | 2.72 | 0.270 |
| 4.0 | 4.45 | 0.440 |
| 6.0 | 5.47 | 0.540 |
| 8.0 | 8.93 | 0.880 |
| 10.0 | 9.24 | 0.910 |

a. $$B = \frac{\sum [PM - \overline{PM}]\ [Pb - \overline{Pb}]}{\sum (Pb - \overline{Pb})^2} = 10.19$$

Where: $\overline{Pb}$ and $\overline{PM}$ are the average concentrations of the species given:

A PM - B Pb

according to the formulas indicated in the same figure.

In this model, it is assumed that the measurement of Pb or any X does not involve any uncertainties, that only the measurements of $PM_{10}$ or Y involve any uncertainties. The uncertainties associated with the model obtained by regression analysis (sometimes termed least squares analysis) can be evaluated by one or more of the statistics that describe the average size of the residuals. The value of $r^2$ (the square of the correlation coefficient) indicates the proportion of the variation in $PM_{10}$ explained by Pb. This is 1.0 for the ideal example of Figure 3.1, and is 0.957 for the data for Figure 3.2. The Standard Error of the Estimate (S.E.E.) is the standard deviation of actual $PM_{10}$ values from those predicted and can be thought of as an "average" residual:

$$\underline{S.E.E} = \frac{(PM-\hat{PM})^2}{N} \qquad \text{Eq. 5.}$$

If the $PM_{10}$ measurements are normally distributed about the least squares line, then approximately 67% of the measured values of $PM_{10}$ for a given Pb concentration will fall within 1 S.E.E. of the predicted value. The uncertainty associated with the estimated value of B can also be estimated and is termed the standard error of B. The statistical significance of the estimated coefficient B can also be tested, usually by evaluating the F ratio. The F ratio is defined as the ratio of the variability in Y explained by the regression line to the unexplained variability in Y, corrected for degrees of freedom. The greater the value of F, the greater the ability of the equation to explain the variability in Y (or in this case $PM_{10}$). The statistical significance of the calculated value of F can be determined by referring to appropriate

statistical tables; however, most computer programs for regression analysis automatically provide this value.

## MULTIPLE REGRESSION ANALYSIS

The simple bivariate linear relationship can be readily extended to the multivariate case in which two or more independent variables, $X_i$, determine the predicted value of Y:

$$Y = A + B_1X_1 + B_2 X_2 + \ldots + B_kX_k \qquad \text{Eq. 6.}$$

For the example discussed in Chapter 2, a model for Pb and V would be of the form:

$$[PM] = A + B_1[Pb] + B_2 [V] \qquad \text{Eq. 7.}$$

This equation implies that both motor vehicle emissions and emissions from oil burning, traced independently by Pb and V, respectively, contribute to the variation in the concentration of $PM_{10}$. When multipled by the ambient concentrations of Pb, the regression coefficient $B_1$ will give the concentration of $PM_{10}$ due to motor vehicles alone. Similarly, $B_2[V]$ gives the concentration of $PM_{10}$ due to oil-burning while A is the concentration of $PM_{10}$ which cannot be explained by variation in Pb or V. The value of A can be due to uncertainties in the measurements of the tracers in the model, or may reflect the presence of unidentified sources of $PM_{10}$ or the inadequacy of the linear model itself. This model, as do other receptor models, assumes that the contributions of $PM_{10}$ from the two sources are linearly additive, a key assumption. Although this assumption appears to be adequate in most instances, further research may ultimately lead to refinements in these models through development

of nonlinear equations.

The regression coefficients in the stepwise multiple regression model are partial regression coefficients. That is, the partial coefficient $B_1$ in equation (7) for example, is equivalent to a simple regression coefficient between $PM_{10}$ and the residuals of [Pb] from which the effect of [V] is removed. That is, if Residual (Pb) = [Pb]-$\hat{[Pb]}$, where [Pb] = A' + B'[V], the partial $B_1$ of equation (7) is equivalent to the simple regression coefficient B.

$$\hat{PM} = A + B\,[\text{Residual (Pb)}] \qquad \text{Eq. 8.}$$

$$PM = A + B_1\,[Pb - [A' + B'(V)]] \qquad \text{Eq. 9.}$$

Thus, partial $B_1$ is the simple regression coefficient between PM and the residuals of [Pb]. The effect of [V] is thus removed from both PM and Pb and the resulting residuals are correlated by simple least squares analysis. As a consequence of this, the coefficient of a given tracer fitted in a stepwise regression will vary somewhat from step to step as each succeeding variable coefficient is fitted.

If there is a great deal of multicollinearity or correlation between the independent tracers, estimates of the regression coefficients may vary considerably from data set to data set. If there is a high degree of correlation between the $X_i$'s, their coefficients may not be uniquely determined or it may not be possible to invert the correlation matrix of these tracer variables. The application of common factor analysis to a data set prior to multiple regression analysis can minimize this problem by identifying statistically independent source tracers to be used as predictor variables. If the source tracers are highly correlated for a given data set, additional field measurements or

additional tracers will be required to separate the contributions of the sources. It should be noted that in some instances, an estimate of the combined contributions of two source types may be adequate for the purpose of a study.

The regression coefficients, $B_i$, relate the source emissions to ambient particulate concentrations according to the relationship:

$$B_i = \frac{\alpha_{ij} [PM]_j}{[T]_j} \qquad j = 1 \text{ to } n \qquad \text{Eq. 10.}$$

where $[PM]_j/[T]_j$ is the ratio of particle mass to tracer mass in emissions from source j and $\alpha_{ij}$ is a coefficient or more complex functional relationship which describes changes in the ratio $[PM]_j/[T]_j$. This may occur between the source and receptor as a result of physical and chemical processes. For the simple example given for Figure 2,

$$B = \frac{1X[PM]_{mv}}{[Pb]_{mv}} \qquad \text{Eq. 11.}$$

where the "mv" subscript denotes motor vehicle emissions and a value of 1 is assumed for $\alpha_{ij}$. With actual air pollution data, the value of $\alpha_{ij}$ need not be 1. If the ratio of $[PM]_j/[T]_j$ for source j emissions has been measured, a value of $\alpha_{ij}$ can be estimated using the regression coefficient from the model for that source and emissions measurements. In principal, more complex relationships could also be determined, although this has not been done to date.

## 3.3 AN EXAMPLE OF AN APPLICATION OF MULTIPLE REGRESSION ANALYSIS TO AIR POLLUTION DATA

Trace element and sulfate concentrations determined for samples collected over a period of two years at an urban site were analyzed by common factor analysis. This solution was shown in Table 2.6. Based on the results, the species Pb, V, Fe, Cu* and $SO_4^=$ were initially selected as predictors for emissions from motor vehicles, oil burning, resuspended soil, incineration and sulfate-related aerosol, respectively, and regression coefficients were fitted for a model of the form:

$$[PM] = A + \sum_{i=1}^{n} B_i [T_i] \qquad \text{Eq. 12.}$$

i = 1 to n sources

Further work indicated that the equation could be improved if Ti rather than Fe was used as a soil tracer. The percent of PM mass contributed by soil resuspension was not significantly changed by this substitution, but the significance of the overall equation was improved somewhat.

Table 3.1 presents the regression coefficients for each of the predictor tracers as well as the estimated standard errors of the coefficients, the values of F and their level of significance. The tracers are listed in the order in which they were entered into the equation by the stepwise regression analysis, i.e., $SO_4^=$ was entered first since it accounted for the greatest proportion of the variability in PM; V was the second predictor fitted, and so on. The coefficients for the first

*Cu could be used as a tracer in this data set as induction motors were used in the air samplers, thus eliminating the usual Cu contamination from the brushes of the Hi-Vol samplers.

Table 3.1. Regression Coefficients for a Multiple Regression Equation for PM[a,b]

| Variable | Source Type | $B_i$ | Standard Error of $B_i$ | Statistics For Coefficients F | P |
|---|---|---|---|---|---|
| $SO_4^=$ | Sulfate-related aerosol | 2.10 | 0.18 | 130.0 | <0.001 |
| V | Oil burning | 110 | 34 | 10.0 | 0.002 |
| Ti | Resuspended soil | 329 | 102 | 10.3 | 0.002 |
| Cu | Incineration | 126 | 64 | 3.8 | 0.05 |
| Pb | Motor Vehicles | 5.8 | 4.4 | 1.7 | 0.19 |
| Constant | - | -1.7 | 3.7 | 0.2 | 0.6 |

[a]Coefficients given for concentrations of tracers expressed in $\mu g/m^3$

$$[PM] = 2.10\,[SO_4^{-2}] + 110\,[V] + 329\,[Ti] + 126\,[Cu] = 5.8\,[Pb] - 1.7$$

[b]Overall statistics for entire equation: $F = 37.3$, $p < 0.001$, $r^2 = 0.72$.

four tracers are all statistically significant at the $p \leq 0.05$ level or better (i.e., there is only a 0.05 probability (out of 1.0)) or less that the relationship between PM and each of the variables is due to chance alone. The coefficient for Pb is of only marginal significance ($p = 0.19$) and would usually be omitted from the equation. However, it can provide an estimate of the motor vehicle contributions at this site, although the estimate will have relatively large uncertainties. The intercept or constant A for this equation is negative but has a large uncertainty and represents only a small fraction of the total PM.

The overall statistics for the entire equation are given at the bottom of Table 3.1. The value of F is statistically significant and the value of $r^2$ (the square of the correlation coefficient) is 0.72, i.e., 72% of the variability in PM is explained by the equation.

At each step in this multiple regression analysis, the tolerance for the variables not yet in the equation was calculated and printed. This can be considered a measure of the independence of the remaining tracers relative to those already entered. A high value (0.8 - 1.0) indicates little association among the tracers (independence) while a low value indicates collinearity with the tracers already in the equation and may result in computational difficulties. In the example given, the tolerances were always greater than 0.8 for tracers not yet in the equation.

Table 3.2 presents the estimates of the average contributions of each type of source to atmospheric concentrations of PM based on the multiple regression coefficients and A of Table 3.1. The average concentration of each of the tracers was multiplied by its corresponding

Table 3.2 Source Contributions to PM Estimated by Regression Equation of Table 3.1.

| Variable | Source Type | Average Concentrations of Tracer $\mu g/m^3$ (Ti) | Regression Coefficients (Bi) | Estimated Source Contribution to PM $\mu g/m^3$[a] (Bi·Ti) |
|---|---|---|---|---|
| $SO_4^=$ | Sulfate-related aerosol | 8.38 | 2.1 ± .2 | 17.6 ± 1.5 |
| V | Oil burning | 0.039 | 110 ± 34 | 4.3 ± 1.4 |
| Ti | Resuspended soil | 0.016 | 329 ± 102 | 5.3 ± 1.6 |
| Cu | Incineration | 0.023 | 126 ± 64 | 2.8 ± 1.5 |
| Pb | Motor Vehicles | 0.506 | 5.8 ± 4.4 | 2.9 ± 2.2 |
| A | Constant | --- | -1.7 ± 3.7 | -1.7 ± 3.7 |

[a]Uncertainties estimated as products of standard error or coefficient times average tracer concentration.

coefficient to obtain these estimates. Sulfate-related aerosol [$(NH_4)_2SO_4$, $(NH_4)HSO_4$, $H_2SO_4$] and any related organic or inorganic species] contributed 17.6 $\mu g/m^3$ of the PM on average, while oil-burning and resuspended soil contributed lower concentrations of 4.3 and 5.3 $\mu g/m^3$, respectively. Incineration and motor vehicles contributed only a few $\mu g/m^3$ at the rooftop (15 stories) sampling site where these measurements were made. Contributions were almost as large as the estimate itself; however, the contributions from this source at the site studied was less than 10% of the total PM.

Once the regression equation has been obtained and examined, it is advisable to examine both the intercept A (PM, which cannot be attributed to known sources) and the residuals of the individual cases. If the value of A is fairly large, it usually indicates that there are unidentified sources of PM. Morandi (2) has shown that the correlation of the residuals of such a model with other variables in a data set can be useful in identifying additional sources of PM. The pattern of the residuals should be examined as well (1). The principal ways of examining such patterns are plots of the distribution of the of the residuals in a time sequence, and plots against both the particulate mass (dependent variable) and the tracers (independent variables) of the model. These techniques, and others, have been discussed in detail by Draper and Smith (1) and can be helpful in arriving at a best possible final regression model for estimating the contributions of various sources.

## 3.4 REGRESSION MODEL VALIDATION

The regression equation obtained in the final stage of the multivariate analysis is a model of the relationship between the dependent variable (in the previous case PM) and a set of independent predictor tracers. The value of such a model lies in its ability to represent the real physical-chemical properties of the system under study, and more particularly of predicting changes or trends likely on the basis of expected changes in source strength. The use of a regression model for these purposes requires validation of the coefficients derived as well as validation of the overall model.

A factor analytical model when satisfactorily validated will provide tracers from which independent predictor tracers will be selected for use in the development of the regression model. Morandi (2) has developed regression subroutines for identification of independent predictor tracers where multicollinearity or variable intercorrelation prevents predictor selection through the factor analysis alone. Regardless of the methodology, the validity of each selected predictor variable must be qualitatively confirmed from published source profiles, known fuel compositions, or well established source emission or source location *vs.* ambient composition relationships. Entry of unexplained predictor tracers in stepwise multiple regression calculations renders outputs of doubtful validity.

During the data analyses the statistical parameters available for evaluation of program outputs should be thoroughly understood and carefully examined at each step in the process. No equation can be considered valid if all statistical criteria have not been met.

The ability of a model to predict the concentration of particulate matter or other air pollutant is a major test of its validity. Adequate tests of predictions are rare in the field of atmospheric pollutant source apportionment. Several approaches to model validation are possible. The Quail Roost II workshop provided a complete set of artificially generated source and receptor composition data for analysis (3). As the true results were known, outputs of the modeling systems were readily validated. Unfortunately, only 40 ambient data sets were generated, which is insufficient for validation of multivariate analysis methods (d/v = 28, see eq. 8, Chapter 2).

Currie, et al. (3) reported the results of the application of several receptor modeling methods to a simlulated set of 40 ambient samples. Factor analysis and multiple regression correctly identified and quantitated a source for which no source profile was provided. One of two chemical mass balance approaches and the Target Transformation (TTFA) method also identified the "missing" source. Another source which produced a simulated 0.05 $\mu g/m^3$ concentration was missed by three methods, and a 20 fold overestimate was reported by a fourth. Two other sources were not found by the regression method; however, one of these sources was identified by TTFA. It appears that the sample set of forty is inadequate to produce accurate FA/MR results.

Individual regression coefficients are the best estimates of the ratio of mass (dependent variable) to the tracer (predictor variable) for a particular source type. For validation these coefficients are readily compared to available source emission profiles. However, care

must be exercised in such comparisons regarding the assumption of the conservative behavior of the emitted particle mass and composition. Alterations of elemental concentrations due to changes in the particle size distribution, condensation, or evaporation enroute to the receptor will affect comparisons of the coefficients with elemental concentrations obtained from stack tests on a particular source.

There are two other approaches that can be used for validation of the regression model. The first is to split the data into a training set and one or more test sets. The training set is analyzed by stepwise multiple regression to obtain the coefficients for the model. The coefficients are then applied to average predictor tracer concentrations to obtain average source contributions. The latter step is performed on both the training and test sets. In each instance, the calculated masses are summed and compared to the measured average for the dependent variable. The comparison affords a measure of the accuracy of the model calculation. This requires the availability of large numbers of samples to produce valid solutions for both data sets.

The ultimate evaluation of a model lies in comparisons of predicted variations with time. Where data bases are collected over sufficient time periods, or ambient studies are repeated, data may be available to compare predictions from early studies to actual results of changes in source strengths or other regulatory activities.

### 3.4.1 Regression coefficient validation

The qualitative validation of the predictor variables has been performed in the examination of factors, factor loadings, clusters of

variables, etc. The stepwise regression produces coefficients (B), which are estimates of the quantitative relationship between each predictor variable and the dependent variable as associated with a particular source (as shown in equation 10). The dependent variable is generally a mass of particles or some substance per cubic meter of air, e.g. B = [Mass in source/V in source].

The coefficients will have error terms which depend on the variance inherent in the data set, but may also have significant biases. This possibility can be examined by validation of the coefficients. Kleinman (4, 5) evaluated pertinent literature data to suggest source profiles for the sources of particles in the New York City area and to validate the derived regression coefficients for several of the major sources. As shown in Table 3.3, the coefficients for 1972-1973 were compared to source emissions composition data to validate the coefficients. The regression coefficients had relative error terms of 25 to 40%. Error terms for the emission ratios cannot be estimated from the available reports.

The case for the automobile, prior to control of the lead content of gasoline, illustrates the ideal since the regression coefficient indicates 12 μg mass/μg Pb. Data for samples taken of intake and exhaust air for a vehicular tunnel were used to calculate a ratio of mass to lead of 11.2 for the automotive emissions added to the exhaust air (7). Automobile exhaust data reported by Mueller *et al.* (11) would have indicated a factor of 2.5. The difference can be attributed to loss of lead by large particle deposition, if any, and mass added to the primary exhaust particles by condensation in the exhaust plume. The

Table 3.3. Comparison of Regression Coefficients and Source Emission Factors

| | μg Particle Mass/μg Tracer | | | | | |
|---|---|---|---|---|---|---|
| | | Literature Emission Ratios | | | | |
| | Regression Coefficient[a.] | Oil Fly Ash[b.] | Auto[c.] | Soil[d.] | Incinerators[e.] | Sulfate Aerosols[f.] |
| V | 103 | 8-91 | | | | |
| Pb | 12 | | 11.2 | | | |
| Mn | 418 | | | 670-4200 | | |
| Cu | 54 | | | | >500 | |
| $SO_4^=$ | 1.66 | | | | | 1.0 to 1.4 |

a. Ambient data from 1972-1973, Kleinman (4, 5).
b. Watson (6).
c. From Larsen, 1966 (7).
d. Watson - for resuspended soil (6).
e. References (8-10).
f. Factors for $H_2SO_4$, $NH_4HSO_4$, and $(NH_4)_2SO_4$.

tunnel study measured the total motor vehicles particulate mass added after dilution and cooling, and gave a coefficient very close to that for the ambient regression coefficient for a period prior to reductions of lead in gasoline or the use of catalyst equipped cars.

The case for vanadium offers a second insight. The concentration of this element in crude oil is highly variable, ranging from less than a few, to hundreds of micrograms per gram of oil, depending on the oil field studied. The concentration in fly ash will vary by source and burner system as well as being dependent on the sampling approach. The agreement for this coefficient was worse, but acceptable.

Profiles for soils and secondary aerosols such as sulfate are not readily obtainable. Both have area or regional sources. The agreements between the regression coefficients for these sources and the available literature data are acceptable, as urban and roadside dusts may contain higher organic contents than reference rocks, lowering any elemental concentrations. Secondary aerosols would typically contain water and organic matter as well as $NH_4^+$.

The copper value falls outside the range of reported ratios for incinerators; major industrial sources such as smelters do not exist in New York City. Either emissions from the residential and commercial incinerators in this area differed from those measured at a few large municipal incinerators studied by others, or an unexpected source was involved.

Daisey and Hopke (12) have compared coefficients from a model for extractable organic matter with organic/elemental masses from source

studies on automotive (Pb), oil burning (V) and resuspended dust sources (Ti). The results showed satisfactory agreement (Table 3.4) particularly in view of the very limited data available in the literature. The variability of the V ratio to the reference values is in part due to the variable vanadium content of oils, which was noted previously.

As part of the Airborne Toxic Element and Organic Substances (ATEOS) (13), a study of sources of inhalable particulate matter in Newark, N.J. by Morandi et al. (14) developed coefficients for sulfate, soil, autos, oil burning and an industrial source. Comparisons to source profile data are shown in Table 3.5.

The ratio of the mass of any emitted substance to a predictor tracer should offer an opportunity for similar comparisons to sources. Daisey (15) and Daisey and Kneip (16) have made such comparisons. Ratios of both extractable organic mass and specific organic compounds to tracers have been evaluated. For the former, a ratio of non-polar extractable organic matter to lead was 0.65 for motor vehicle sources on a annual basis and 2.2 with no space heating. This compares to an estimate of 1.6 for Allegheny Tunnel data. Daisey (15) also found that a regression coefficient for chrysene/Pb of 0.0017 compared well with an emission estimate of 0.0022.

Morandi (2) has continued the evaluation and validation of predictor variable coefficients by methods which emphasize the examination of the intercept and the residuals of a regression equation for inhalable particulate matter. A pattern in the time sequence of the inhalable particulate matter residuals indicated a missing predictor tracer.

Table 3.4. Comparison of Source Tracer Coefficients and Ratios of Non-Polar Cyclohexane-Soluble Organics (CX) or Volatilizable Carbon to Tracer in Source Emissions

| | Organic Species | µg Extracted Mass/µg Tracer Coefficient of or Ratio to: Pb | V | Ti | Ref. |
|---|---|---|---|---|---|
| Models | | | | | |
| Model I(1979,1980) | CX | 1.1 ±0.4 | 52 ± 3 | 20±12 | 12 |
| Previous NYC Model(1978) | CX | 0.65±0.35 | 29 ± 4 | - | 16 |
| Sources | | | | | |
| Allegheny Tunnel,1979[a]. Spark ignition engines | CX | 2 ±7 | | | |
| Residual Oil(fine particle) | Volatilizable Carbon | | 5.7±34.0<br>0.6- 7.8<br>0.1-10.3 | | 7 |
| Street Dust(fine particle) | Volatilizable Carbon | | | 11-13 | 7 |

a. Samples of particulate matter collected in the Allegheny Tunnel during the summer of 1979 were provided by Dr. William Pierson of the Ford Motor Company. Portions of each sample were extracted sequentially with cyclohexane and dichloromethane to determine extractable organic masses. A second aliquot of each was analyzed for lead.

Table 3.5. Comparisons of Regression Coefficients of Particulate Matter Models to Source Emissions Compositions

| | μg Mass/μg Tracer | | | | | |
|---|---|---|---|---|---|---|
| | | | Literature Emission Ratios | | | |
| Predictor Variable | Regression Coefficients | | Sulfate Aerosols[a.] | Soils[a.] | Autos[b.] | Oil[a.] |
| | Newark | New York[a.] | | | | |
| $SO_4^=$ | 1.6 | 1.66 | 1.0-1.4 | | | |
| Mn | 718 | 418 | | 670-4200 | | |
| CO[b.] | 714 | - | | | 350-700 | |
| V | 106 | 103 | | | | 8-91 |

a. See Table 3.3.
b. Substituted for Pb as multiple lead sources were observed in the factor analysis. Emissions estimates from Allegheny Tunnel Studies (17, 18).

The analysis of the regression equation residuals has enabled Morandi (2) to develop predictor tracers based on tracer regressions and analysis of residuals to identify a tracer for an unsuspected source. Meteorological relationships, to be discussed, and micro inventory processes, discussed earlier, have aided in the validation of the residuals from the tracer regression equations.

### 3.4.2 Meteorological Relationships

Studies of point sources have, for obvious reasons, often emphasized the changes in concentrations associated with samples taken simultaneously both upwind and downwind of such a source. The data obtained in a source apportionment study is often from a temporal (time series) study design rather than a spatial design. Spatial information is, however, embedded in the sample data because of changes in wind direction. Morandi et al. (14) have found that specific time periods or single days may yield extremely high concentration values for one or more tracers which are not characteristic of the data distribution found for the bulk of the data. In such cases careful use of data truncating methods may be required to obtain a useful set of factor analysis results or to reduce apparent error in the regression coefficients (19).

Several common sources of error or causes of tracer intercorrelation include meteorological relationships, and sampling and analytical influences. The latter two problem types should be examined so that invalid data are removed before multivariate statistical analysis procedures are applied. As this is not always done effectively, reviews are necessary where unusual outcomes are observed (20, 21).

More commonly, specific data sets with extreme values may relate to unusual meteorological conditions such as stagnations, repeated daily inversions, or continuous stable wind directions. These periods, whether single days or sets of days often termed "episodes", should be studied to determine the relationship of the unusual predictor tracer concentrations to meteorological factors. The regression coefficients obtained with reduced data sets must also be compared to those for total data sets, and, if possible, with the excluded data sets. This latter step forms a further basis for coefficient validation.

Examination of the extreme value sets was actually used in the previously cited example for Newark (2). High zinc concentrations occurred during an episode and were strongly related to northeasterly winds. This wind direction occurred ≈20% of the time for the Newark area, and proved to be the direction which put the sampler downwind of the smelter.

3.4.3 General Model Evaluation

The regression equation is a numerical model of the source-receptor relationships which were the focus of the program. The overall validation of the equation requires more than validation of the individual coefficients or examination of source-receptor geography and meteorological parameters.

The model must be examined as a whole. the residual term must be satisfactorily small. An objective means of determining this requires setting a limit prior to the study. A residual term of more than 10% of the total measured dependent variable maybe an indication that a

significant source has been missed. The presence of predictor variables expected for local sources, and general tracer relationships compatible with known sources provided further confidence in the overall model. Should an expected major source fail to have a significant coefficient for an acceptable tracer, serious doubt is cast on the model. Overall knowledge of the area source inventory and catalogs of source emission composition profiles are useful in these types of general model evaluations. The model must have an overall sound, logical relation to sources, meterology and atmospheric chemistry and physics.

Additional objective evaluations are possible in model validations where sufficient data are available. For instance, a model may be based on data called a "training set" and then used with additional data called a "test set." In the process of the development of the Factor Analysis/Multiple Regression approach a long term data base was collected for several locations in New York City. Kleinman (4) and Kleinman *et al.* (22) demonstrated the validity of the approach by several evaluations of "test sets," i.e., data sets unused in the regression analysis, using coefficients for the 1972-1973 data for samples taken at the New York University Medical Center, i.e., the "training set."

The coefficients obtained for the 1972-1973 data were first used to calculate particle mass for each source and the total mass. The residual term was also recorded for that period. The calculated sum was compared to the average measured TSP as a means of validating the model. As would be expected, the percentage difference between calculated and observed TSP was only -3.7% for 1972 and +5% for 1973.

The regression coefficients were adjusted for known changes in fuel composition and the adjusted coefficients were used with the respective yearly average predictor variable concentrations for 1969, 1974 and 1975 for the same sampling site. The differences were +9.7%, +7% and +2.8% for measured TSP values of 134 $\mu g/m^3$ (1969), 71 $\mu g/m^3$ (1974) and 52 $\mu g/m^3$ (1975). While the model appears to accurately identify the major sources, it apparently failed to indicate correctly the decreased emissions when annual TSP levels dropped from 70 to 80 $\mu g/m^3$ in the 1972 to 1974 period to 52 $\mu g/m^3$ in 1975. The residual term accounted for a third or more of the calculated mass in the 1975 calculation. Therefore, possible changes in the regression coefficients, i.e., emission composition changes, and variation in the residual term may have seriously affected the accuracy of the estimated TSP values for the test set for data collected in 1975.

The coefficients derived from the 1972-1973 results for the Medical Center site were also applied to data for sites in lower Manhattan, the Bronx, Queens and a rural site in Tuxedo, N.Y. The differences between calculated and observed values ranged from 0 to 13% for the first two sites which are affected by sources similar to those around the central Manhattan Medical Center site. For the Queens and the rural area differences of -32 to 57% were observed indicating that the model could not be used at these two sites; this is probably due to substantial differences in the nature of the sources at these two sites compared to the Manhattan site.

In a subsequent study Kneip, Mallon and Kleinman (23) observed variations year to year in the coefficients over the period of 1977 to

1980. The shifts were probably related to changing fuel compositions and changes in sources. Satisfactory differences between calculated and observed values were obtained for the 1978 to 1980 data when respirable particles ($D_{50} \leq 3.5$ μm) and coarse particles ($D_{50} \geq 3.5$ μm) were sampled and the tracers were determined for each size fraction. The coefficients for each size range gave measured differences of -2% for the respirable fraction and 0% for the coarse fraction.

They concluded that regression coefficients were valid for the periods during which data was collected, but not valid for other periods where unknown or undefined changes in fuels or sources had affected the source emission compositions. For instance, a shift in the coefficient for vanadium was apparently related to a sharp increase in the airborne vanadium levels. A change in the location of the sources of crude oil caused by the Iranian revolution probably affected the source emission compositions and ambient concentrations. Copper concentrations dropped sharply from 1976 to 1980 as small residential incinerators were shut down in New York. The copper coefficient was no longer significant in the 1978-1980 data.

### 3.4.4 Summary

The validation process is undertaken in these stages, 1) validation of the source related factors obtained and predictor tracer selected; 2) validation of the regression coefficients; and 3) overall model validation. While these have been successfully performed in a number of cited cases, several principles must be kept in mind. Sufficient data sets must be available to obtain a reasonable number of significant factors. Under or over specification in the selection of probable predictor

tracers can be a problem. The presence of only a few factors or a large residual term in the regression equation usually signifies a lack of sufficient information on potential sources and the related predictor tracers. Careful experimental designs and validation efforts not only will provide greater confidence in the factor-source and quantitative regression relationships, but also will aid in defining sources not expected when the original design was established.

## 3.5 INTERPRETATION

The final interpretation of a source apportionment/receptor model is dependent on all preceding steps. Clearly stated objectives, effective experimental design, accurate and precise operations from procurement to data base maintenance, thorough multivariate analyses, and complete model validation are all essential. Since the FA/MR method is applicable in situations of standard compliance problems or toxic substance investigations, it is assumed that a major objective is the definition of those sources which may be controllable, and may also be contributing significant fractions of the pollutant under study. The final product of such a study is the concentration of pollutant assignable to each source at the receptor during the period of the study. Fractional contributions are normally calculated and the magnitude of each source contribution is then apparent.

The first step in interpretation is a comparison of all data for the pollutant in question to a standard or to exposure limits. An estimate of overall reductions needed is made on the basis of this comparison. The second step is to examine the source assignments and the

residual term from the regression equation. If the latter is a large fraction of the total pollutant concentration, a major source or sources may remain unidentified. When this occurs, one returns to the factor analysis step and reexamines the assumptions concerning convergence, communality and significance of the apparently less important factors (2).

Assuming this is not the case, the sources identified must be categorized as controllable, potentially controllable or uncontrollable. The latter might be a case for the problem of regional secondary sulfate pollution. The situation may remain virtually unchanged, since any controls implemented to reduce local contributions to zero will not reduce the concentrations at the receptor site from the regional sources.

Controllable or potentially controllable sources are the categories desired when major reductions in ambient concentrations are needed. The previous sections have illustrated cases of industrial point sources, area sources, automotive emissions, and general industrial contributions. Each of these or any other source discovered must be reviewed prior to establishing final conclusions.

The criteria are for example:

1. Type of source

   Point - major

   Point - numerous, small

   Area - resuspended soil

Mobile - motor vehicles

2. Fraction of mass attributable to each source

3. Error range of mass

4. Validity of factor and predictor variable for each source

5. Validity of each regression coefficient

6. Meteorological relationships

Wind direction

Stability

Stagnation episodes

7. Micro and Macro Inventories

8. Observation of source operations.

Several decisions can be made on the basis of these criteria. A source (point or area) may be judged to be a major contributor, and controlled. Thus, a control program should be established. A source(s) could be judged insignificant or uncontrollable and no control program implemented. A source(s) could be judged significant, but more information needed to clearly prove the case.

Such a case might occur where the FA/MR method identifies a previously unknown, unevaluated or unexpected source. These may include such cases as: groups of small commercial/industrial operations with poor emissions controls or frequent outdoor operations not controlled in any

way; a large point source impacting the receptor via infrequent wind directions or only during extended stagnation episodes; or impacts of resuspended soil from vacant lands. These types of problems may require special studies where the data are judged insufficient for action.

### 3.5.1 Summary

Interpretation of the FA/MR results is based on a well designed program, operated and evaluated by well informed staff members.

It is important to maintain an appreciation of the relationship between objectives and interpretation, to carefully document every step of the process and to use all information from the validation steps in classifying the results of the overall interpretation.

## 3.6 References

1. Draper, N., and Smith, H., 1981, Applied Regression Analysis, 2nd ed., John Wiley and Sons, New York.

2. Morandi, M.T., 1985, Development of receptor oriented source apportionment models for inhalable particulate matter and particulate organic matter in New Jersey. Ph.D. Dissertation, New York University Medical Center, February.

3. Currie, L., Gerlach, R.W., Lewis, C., et al., 1984, Intercomparison of source apportionment procedures: Results for simulated data sets. Atmos. Environ., 18: 1517-1537.

4. Kleinman, M.T., 1977, The apportionment of sources of airborne particulate matter. Ph.D. Dissertation, New York University Medical Center, June.

5. Kleinman, M.T., Eisenbud, M., Lippmann, M., Kneip, T.J., 1980, The use of tracers to identify sources of airborne particles. Environ. Int., 4: 53-62.

6. Watson, J.G., 1979, Chemical element balance receptor model methodology for assessing the source of fine and total suspended particulate matter in Portland, Oregon. Ph.D. Dissertation, Oregon Graduate Center, February.

7. Larsen, R.J., 1966, Air pollution from motor vehicles. Ann. N.Y. Acad. Sci., 136: 275.

8. Greenberg, R.R., Gordon, G.E., Zoller, W.H., Jacko, R.B., Neuendorf, D.W., and Yost, K.J., 1978, Composition of particles emitted from the nicosia municipal incinerator. Env. Sci. Tech., 12(12): 1329-1332.

9. Law, S.L., and Gordon, G.E., 1979, Sources of metals in municipal incinerator emissions. Env. Sci. Tech., 13(4): 432-438.

10. Jacko, R.B., and Neuendorf, D.W., 1977, Trace metal particulate emission test results from a number of industrial and municipal point sources. J. APCA, 27(10): 989-994.

11. Mueller, P.K., Helwig, H.L., Aleocer, A.E., Gong, W.K., and Jones, E.E., 1964, Concentration of fine particles and lead in car exhaust. Symposium on Air Pollution Measurement Methods, Special Technical Publication, No. 352, American Society for Testing Materials.

12. Daisey, J.M., and Hopke, P.K., 1984, Receptor source apportionment models for three fractions of respirable particulate organic matter. Paper No. 84-16.4 presented at the 77th Annual Meeting of the Air Pollution Control Association, San Francisco, CA, June 24-29.

13. Lioy, P.J., Daisey, J.M., Atherholt, T., Bozzelli, J., Darack, F., Fisher, R., Greenberg, A., Harkov, R., Kebbekus, B., Kneip, T.J., Louis, J., McGarrity, G., McGeorge, L., and Reiss, N.M., 1983, The New Jersey Project on airborne toxic elements and organic substances (ATEOS): A summary of the 1981 summer and 1982 winter studies. J. APCA, 33(7): 649-657.

14. Morandi, M., Daisey, J.M., and Lioy, P.J., 1983, A receptor source apportionment model for inhalable particulate matter in Newark, N.J. Paper No. 83-14.2 presented at the 76th Annual Meeting of the Air Pollution Control Association, Atlanta, GA, June 19-24.

15. Daisey, J.M., In Press, A new approach to the identification of sources of airborne mutagens and carcinogens: Receptor source apportionment modeling. Env. Int'l.

16. Daisey, J.m. and Kneip, T.J., 1981, Atmospheric particulate organic matter - Multivariate models for identifying sources and estimating their contributions to the ambient aerosol. ACS Symposium Series, No. 167, Atmospheric Aerosol: Source/Air Quality Relationships, E.S. Macias and P.K. Hopke, Eds., American Chemical Society. 197-221.

17. Pierson, W.R., and Brachaczek, W.W., 1983, Particulate matter associated with vehicles on the road, II. Environ. Sci. Tech., 2: 1.

18. Gorse, R.A. and Norbeck, J.M., 1981, CO emission rates for in use gasoline and diesel vehicles. J APCA, 31: 1094.

19. Shah, J.J., Huntzicker, J.J., Kneip, T.J., and Daisey, J.M., 1981, Investigation of the sources of carbonaceous aerosol in New York City by multiple linear regression. Paper No. 81-64.3 presented at the 74th Annual Meeting of the Air Pollution Control Association, Philadelphia, PA, June 21-26.

20. Gaarenstroom, P.D., Perone, S.P., Moyers, J.L., 1977, Application of pattern recognition and factor analysis for characterization of atmospheric particulate composition in southwest desert atmosphere. Env. Sci. Tech., 11: 795-800.

21. Daisey, J.M., Licy, P.J., and Kneip, T.J., 1984, Receptor models for airborne organic species. EPA report submitted January, Contract No. CR810300-01-0.

22. Kleinman, M.T., Pasternack, Eisenbud, M., and Kneip, T.J., 1980, Identifying and estimating the relative importance of sources of airborne particulates. Env. Sci. Tech., 14: 62-65.

23. Kneip, T.J., Mallon, R.P., Kleinman, M.T., 1983, The impact of changing air quality on multiple regression models for coarse and fine particle fractions. Atmos. Env., 17(2): 299-304.

# APPENDIX

## 4.0 ALTERNATIVE APPROACHES TO REGRESSION ANALYSIS

Several authors have used variations in the factor analysis steps to aid in selection of the independent predictor variables, i.e. tracers, used in the stepwise multiple regression. Three methods reported to date are described briefly in this appendix.

### A.1 TARGET TRANSFORMATION FACTOR ANALYSIS (TTFA)

Both principal component and common (classical) factor analysis apportion the variance rather than the absolute concentration of the elemental tracers of a data set among different factors. As a consequence, the source-related factors which are obtained do not give the relative concentrations of the elements in the emissions from each source nor the contributions of each source factor to the ambient particle concentrations. In order to overcome these inherent limitations, Hopke and co-workers (1-4) have developed the target transformation factor analysis (TTFA) for source apportionment of the ambient aerosol. The TTFA method differs from PCA and common factor analyses in two important ways:

1) the use of Q rather than R-mode analyses;

2) rotation of the obtained factors to align with a target which is a source emission profile rather than rotation according to abstract statistical criteria.

As the first step in TTFA, Q-mode factor analysis is used to screen the data and to determine the upper limit on the number of factors to be

retained for the actual TTFA. In Q-mode analysis, correlations between samples rather than between variables (i.e., elements) are used and the relative elemental concentrations are thus preserved in the analysis (2). For example, ambient Pb concentrations are typically one to two orders of magnitude greater than those of Cd and, in Q-mode analysis, this feature of the data is retained. In the more usual R mode analysis, the data for each element are normalized in calculating the correlations between elements and the information about relative elemental concentrations is thus lost. In Q-mode analysis, the source profile matrix is calculated first from the data and eigenvalue matrix; in R-mode, the mass contribution matrix is calculated first from the data and eigenvectors. In principal, the results of the two analyses should be comparable and this is generally observed in practice.

The second difference between TTFA and other types of factor analyses lies in the method of rotating the factors in order to obtain interpretable factors. In TTFA, each factor is rotated to maximize (by least squares analysis) the overlap of the factor and a test vector. The test vector can be an actual source emissions profile containing values for all of the elements (normalized relative to one element in the vector) or an unique test vector in which a tracer element is assigned an initial value of 1.0 and all other elements are assigned a value of zero. The latter is the most common approach. When unique test vectors are used, there is a vector for each element or species. The test vector is iteratively refined by replacing the input value for a given element with that predicted by the target transformation and the factor is again rotated against the refined test vector. This should be equivalent to refining source emissions profiles measured at a source to

the form in which they exist in the atmosphere.

After each of the refined source profiles has been obtained they are normalized to a sum of 1,000,000 and they are grouped p vectors at a time to reproduce the original ambient data. Cluster analysis is used to determine which of the iterated profiles are similar and these are combined to form a matrix of source profiles. These source emissions profiles are then scaled to reflect the actual concentration of each element in the emissions from that source (μg element per gram of particles emitted) by regression of the mass concentration of each sample against each element. Finally the contribution of each source to total particle concentration is determined based on the scaled source emissions profiles.

The FANTASIA program written by Hopke and co-workers (1-4) to perform TTFA is available through the courtesy of Dr. Hopke. The program has been written for the CDC Cyber 175 Computer and utilizes a number of subroutines from the IMSL subroutine library (5) for matrix manipulations. The principal advantage to TTFA is that complete source emissions profiles can be obtained from ambient measurements (based on use of unique vectors) and can be compared to actual emissions profiles, thus providing evidence of the validity of the source apportionment model.

## A.2 MULTIPLE LINEAR REGRESSION ON TRACERS/FACTOR ANALYSIS [MLR(T)/FA]

Dattner (in Currie, 6) has applied the FA/MR method to the data sets analyzed in the Quail Roost II exercises. This version of the method was designated 'Multiple Linear Regression by Tracers/Factor

Analysis [MLR(T)/FA]. The steps described were classical factor analysis for data screening and for determination of the number and characteristics of major sources, selection of the element with highest loading as a tracer for each factor. "Backward stepwise unweighted multiple linear regression" was used to obtain the equation for mass in terms of the tracer elements with significant regression coefficients (coefficients, $B_j > 1.96\ S_{B_j}$) and calculation of the source contributions from the products of the coefficients, $B_j$ and concentrations, $C_j$. The source contributions were calculated for each sampling period (case) and averaged over cases of interest.

Dattner extended the method of repeating the backward stepwise regression for each tracer as dependent variable, and of normalizing the regression coefficient matrix obtained to a total mass coefficient for each source (column) equal to unity. The columns are then concentration equivalent profiles which can be compared to available source profile data sets.

## A.3 REGRESSION OF ABSOLUTE PRINCIPAL COMPONENT SCORES (1)

Principal Component Analysis is applied to elemental and other composition variables for a large data set, in the case of Thurston and Spengler (2) 332, and the results are used to identify the sources affecting the site. Then the mass contribution from each source is identified using a new empirical technique that involves computation of an Absolute Principal Component Score (APCS) for each sample, and regression of the sample mass on the APCS to obtain a mass contribution. Pollution source elemental profiles suggested by the source impacts of

the final regression analyses are then compared with the literature.

The principal component analysis used is the standard type explained in Chapter 2, with a varimax rotation applied to obtain the best orthogronal representation of the factors. After examination of the original composition variable loadings on a given factor, the results are interpreted and identifiable sources noted.

The next step in the analysis is the examination of the principal component scores (eq. 3, Chapter 2) to estimate the quantitative source impacts. Thurston and Spengler (2) note that the principal component scores are correlated with a source impact but are not proportional to these impacts in the usual Z-score of PCA computer printout. Regression of the dependent variable on these scores ($P_{jk}$) would be of the form

$$Y_k = \overline{Y}_k + \sum_{j=1}^{p} \zeta_j P_{jk} \qquad (7)$$

where $\overline{Y}_k$ is the mean, and $\zeta_j$ are the conversion coefficients of the non-dimensional score deviations into mass deviations from the mean impact of a source (7). Since these results are not deviations from zero, there is no apportionment possible of the dependent variable (e.g. TSP, IPM, PM).

The technique developed by Thurston and Spengler (2) estimates the absolute zero principal component score by separately scoring an extra "day" in which the tracer concentrations were zero. This is accomplished by obtaining a Z-score (Chapter 2, eq. 2) for its absolute zero concentrations, and then calculating the rotated absolute zero PC

scores. These estimated absolute zero scores are subtracted from the original components for each day to obtain an APCS.

The final set in the solution to obtain the mass contributing from a given source is

$$M_x = \gamma_o + \sum_{j=1}^{p} \gamma_j \, APCS^*_{jk}$$

where $M_k$ is the mass associated with observation k, $APCS^*_{jk}$ is the rotated absolute component scores for source j and observation k, and $\gamma_j APCS^*_{jk}$ is the mass contribution from source k.

## A.4 REFERENCES

1. Hopke, P. K., R. E. Lamb and d. F. S. Natusch, 1980. Multielemental characterization of urban street dust, Environ. Sci. Technol., 14: 164-172.

2. Alpert, D. J. and P. K. Hopke, 1980. A quantitative determination of sources in the Boston urban aerosol, Atmos. Environ., 14: 1137-1146.

3. Alpert, D. J. and P. K. Hopke, 1981. A determination of the sources of airborne particles collected during the regional air pollution study, Atmos. Environ., 15: 675-687.

4. Hopke, P. K., D. J. Alpert and B. A. Roscoe, 1983. Fantasis - A program for target transformation factor analysis to apportion sources in environmental samples, Computers an Chemistry, 7: 149-155.

5. IMSL Library Reference Manual, IMSL LIB-0008, IMSL, Inc., Houston, Texas.

6. Currie, L. A. et al., 1984. Intercomparison of source apportionment procedures: results for simulated data sets, Atmos. Environ., 18: 1517-1537.

7. Thurston, G. D., 1983. A source apportionment of particulate air pollution in metropolitan Boston. Doctoral Dissertation. Department of Environmental Health Sciences, School of Public Health, Harvard University, Boston, MA.

8. Thurston, G. D. and J. D. Spengler 3, 1985. A quantitative assessment of source contributions to inhalable particulate matter in metropolitan Boston, Atmos. Environ., 19: